REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author Brooks, S.P.

Molecular genetic and functional analyses of X-linked congenital cataract

Simon Phillip Brooks BSc (Hons)

A thesis submitted for the Degree of
Doctor of Philosophy

Department of Molecular Genetics
Institute of Ophthalmology
University College London
Bath Street
London
EC1V 9EL

Institute of Ophthalmology

2006

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I declare that this thesis submitted for the Degree of Doctor of Philosophy is composed by myself, and the work herein is my own, or that the author involved is clearly stated.

Simon Phillip Brooks

Abstract

Nance-Horan Syndrome (NHS) is an X-linked developmental syndrome characterised by congenital cataract, dental anomalies, and dysmorphological features often associated with mental retardation. The NHS locus on Xp22.13 is encompassed by the disease locus for X-linked congenital cataract (CXN). Analysis of microsatellites within the CXN family resulted in refinement of the CXN disease interval, reducing the region of overlap between the CXN and NHS disease loci. Candidate genes in the overlapping intervals were identified bioinformatically and their genomic structures evaluated. Patient DNA was screened by direct sequencing, resulting in the identification of mutations within a novel gene in four British families with NHS, but not the CXN family. This novel gene, named *NHS*, is encoded by at least 10 exons transcribed into at least five mRNA isoforms A, B, C, D, and E (encoding a putative 1,630 a.a., 1,335 a.a., 1,474 a.a., 1,453 a.a., and 1,473 a.a. protein, respectively). All mutations identified are truncating and three mutations have been identified in exon 1, which are only expressed in isoform A. This implies that mutations in isoform A are sufficient to cause disease in families with NHS. Functional clues for the NHS protein were investigated resulting in identification of three new genes with significant homology to *NHS* (*NHS-Like 1* (*NHSL1*), *NHSL2* and *NHSL3*). All four genes share a conserved genomic structure. Fetal expression analysis of *NHS*, *NHSL1* and *NHSL2* suggests that *NHSL1* and *NHSL2* are more ubiquitous than *NHS*. Analysis of the NHS family of proteins revealed significant homology to members of the WASP family, which consists of WASP, N-WASP and WAVE1-3. The WASP protein family play a crucial role in regulating actin dynamics, directly linking small GTPase signalling to actin polymerisation through activation of the Arp2/3 complex. An anti-peptide antibody to the C-terminus of NHS, completely conserved across species, was raised and characterised. A major NHS isoform (approximately 170 kDa) was detected in several cell lines. Subcellular localisation studies in MTLn3 cells showed localization of endogenous NHS to the leading edge of lamellipodia, a localisation pattern reminiscent of the Arp2/3 complex. Endogenous

NHS also localised to some actin stress fibres. Homology to the WASP protein family and localisation of endogenous NHS to the leading edge of lamellipods strongly supports a role for NHS in actin cytoskeletal dynamics during development.

Acknowledgements

To say thanks in an acknowledgements section is by no means enough to express how much I appreciate everything Alison Hardcastle (my primary supervisor) has done for me. You didn't just provide me with the opportunity to undertake a Ph.D., which has exceeded all my hopes and expectations, but you gave me an opportunity to work in what I can only describe as an amazing group. My Ph.D. has been so enjoyable and interesting and you have given me so much support that I would do it all over again! Thanks a million times over to Jane, Neil, and Suba. Working with you all has been brilliant and I've much appreciated all your support and advice but more importantly your friendships.

I owe a huge thanks to Michael Cheetham for all your advice and teachings, especially on the latter half of my Ph.D., and for giving up so much of your time. You too have a great group who have been very supportive and a lot of fun socially. So a huge thanks to Hugo, Jaqui, Juan, Paul, Naheed, and your newest members Maria, Sergey, and Tatyana.

I would like to thank Maryse Bailly for your time and advice on the functional side of my work and for providing some valuable resources. I look forward to our future collaborations. Thanks to Dongmin Shao for help with the actin polymerization assay. Huge thanks to Beverley Scott for ensuring that the sequencer ran smoothly. Thank you to Peter Munro for help with using the confocal microscope. Thanks to Isabelle Russell-Eggitt and Tony Moore, and to all those at the Institute who have made it a great place to work at. A special thanks to Louise, Merv, Naheed R, Resh, Sakina, and Shazeen.

Thanks to Fight for Sight, especially the Wiltshire committee who raised all the funds for this PhD. It was a pleasure to meet you.

Sincere thanks to my parents for their support throughout, without all your encouragement I most definitely wouldn't be where I am now. You can now take a sigh of relief that at 26, I shall no longer be a student! Thanks to my "little" brother and Danielle for being there and to Kathy and Allan who have always been very supportive. Last but not least, I would like to thank my Caroline. You don't realise how much you have done for me over the years. You are by far, one in a zillion.

Dedication

I dedicate this thesis to Caroline who has been so supportive throughout, and to my family and friends.

Table of contents

CHAPTER SIX

List of figures

List of tables

Abbreviations

Abi	Abelson-interactor 1
APS	Ammonium persulphate
Arp2/3	Actin related protein 2 and 3
ATP	Adenosine 5'-triphosphate
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search tool
bp	Base pair
BSA	Bovine serum albumin
Caco2	Human Caucasian colon adenocarcinoma
cDNA	Complementary deoxyribonucleic acid
CHO	Chinese hamster ovary
COS7	Monkey African Green Kidney
CXN	X-linked congenital cataract
DAPI	4'-6 Diamidino-2-phenylindole
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonuclei acid triphosphate
ECL	Enhanced chemiluminescence
EST	Expressed sequence tag
ExPASy	Expert Protein Analysis System
GBD	GTPase binding domain
GST	Glutathione-S-transferase
HRP	Horse-radish peroxidase
IAA	Isoamylalcohol
IgG	Immunoglobulin G
IPTG	Isopropyl-thiogalactoside
kDa	Kilo Daltons
kb	Kilo base
MDCK	Dog cocker spaniel kidney
mRNA	Messenger ribonucleic acid

MTLn3	Metastatic rat mammary adenocarcinoma
N-WASP	Neural-WASP
NaOAc	Sodium acetate
NaOH	Sodium Hydroxide
NCBI	National Center for Biotechnology Information
NCG1	Novel candidate gene 1
NHS	Nance-Horan Syndrome
MIM	Mendelian Inheritance in Man
PAGE	Polyacrylamide gel electrophoresis
PAC	P1 derived artificial chromosome
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PIP2	Phosphatidylinositol 4,5-bisphosphate
PtdIns(3,4,5)P_3	Phosphatidylinositol 3,4,5-trisphosphate
RNA	Ribonucleic acid
RT-PCR	Reverse-transcription polymerase chain reaction
SDS	Sodium dodecyl sulphate
SDW	Sterile distilled water
SHD	Scar Homology Domain (a.k.a. WHD)
SGP	Syntenic Gene Prediction
SNP	Single nucleotide polymorphism
SOC	Super optimal catabolite
TAE	Tris acetate EDTA
TEMED	N, N, N', N'-Tetramethylethylenediamine
Tris-HCl	Trizma hydrochloride
v/v	Volume/volume
w/v	Weight/volume
WASP	Wiskott-Aldrich Syndrome protein
WAVE	WASP-family verprolin homologous protein
WCA	WASP homology 2/Connecting region/Acidic region
WH1	WASP homology 1
WH2	WASP homology 2
WHD	WAVE Homology Domain (a.k.a. SHD)
Xcat	X-linked congenital cataract of the mouse

- Chapter One -

Introduction

1.1 Structure of the eye

The human eye consists of three basic layers, a fibrous (corneoscleral) coat, the uvea or uveal tract (formed from the choroid, ciliary body and iris) and the neural layer (the retina). As shown in figure 1.1, the human eye contains numerous structures. The pupil forms the central aperture of the iris, a pigmented circular muscle. The iris controls the amount of light passing through the lens by controlling the size of the pupil. Due to the absorbing pigments of the retinal pigment epithelial cells in the retina, the pupil appears black. The cornea, forming the fibrous coat, is responsible for most of the refractive power of the eye. The cornea is transparent, covering the lens and the iris, and presents a tough physical barrier to trauma and infection. The sclera, which is continuous with the cornea, is a dense fibrous opaque white outer coat, forming part of the fibrous coat.

The lens (comprised of the capsule, lens epithelium and the lens fibres) is suspended in the eye by zonular ligaments between the aqueous humour and the vitreous body. The zonular ligaments cause the lens to change shape by contracting and relaxing (known as accommodation), thereby allowing a sharp image to be formed on the retina.

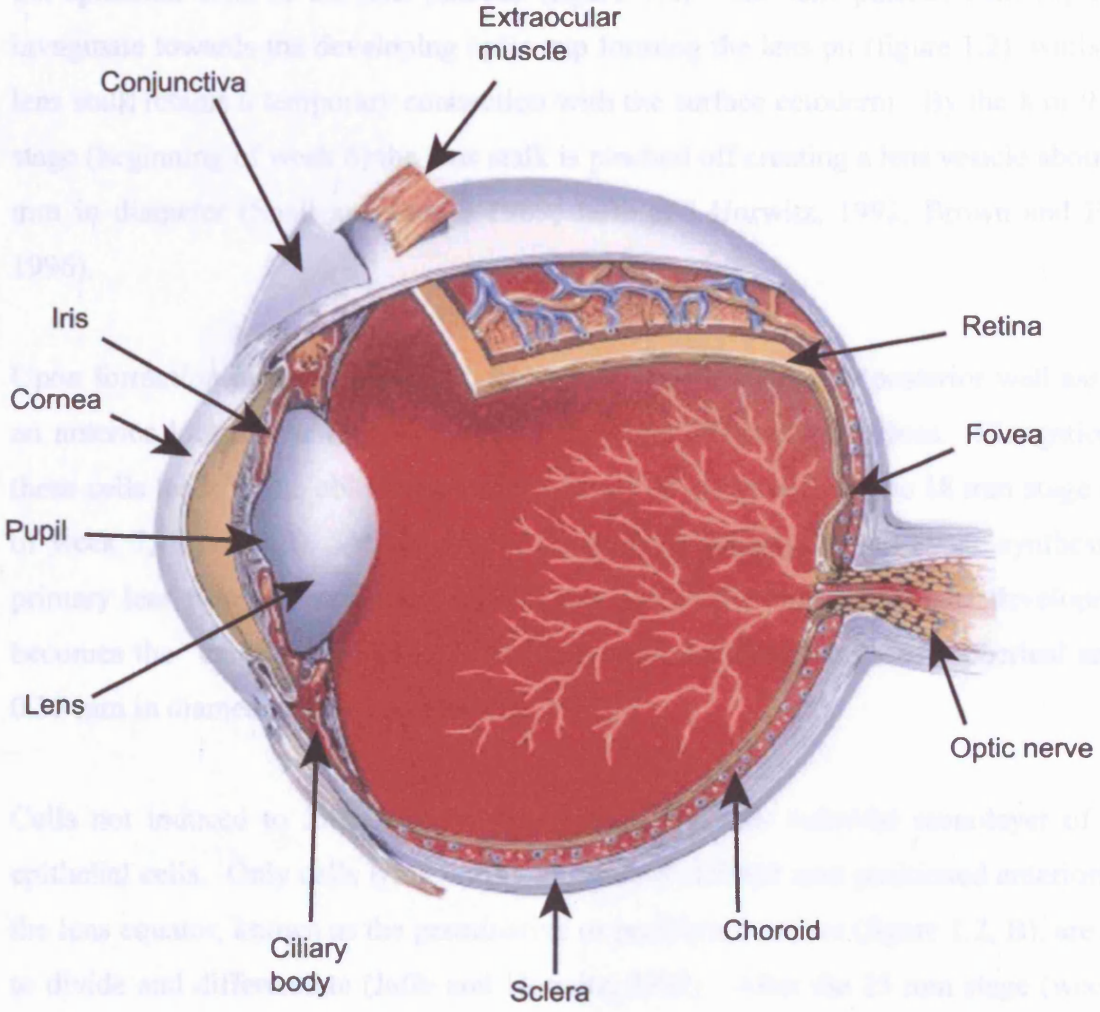


Figure 1.1 A cross-sectional view of the human eye
(Adapted from Fight for Sight, http://www.fightforsight.org.uk/)

1.2 Lens development and cataract morphology

1.2.1 Lens development

Human lens organogenesis can be detected at 3-4 weeks gestation (4-mm embryo). Surface ectoderm cells overlying the developing optic vesicle starts thickening to form the epithelial cells of the lens placode (figure 1.2). The lens placode then begins to invaginate towards the developing optic cup forming the lens pit (figure 1.2), whilst the lens stalk retains a temporary connection with the surface ectoderm. By the 8 or 9 mm stage (beginning of week 6) the lens stalk is pinched off creating a lens vesicle about 0.2 mm in diameter (Snell and Lemp, 1989; Jaffe and Horwitz, 1992; Brown and Bron, 1996).

Upon formation of the lens vesicle, the nuclei of cells lining the posterior wall assume an anterior location as they elongate to form the primary lens fibres. Elongation of these cells leads to the obliteration of the cavity of the vesicle by the 18 mm stage (end of week 7, figure 1.2). At the end of the seventh week of fetal life the synthesis of primary lens fibres is complete, and it is this part of the lens that in later development becomes the "embryonic nucleus." By this stage the lens is primarily spherical and is 0.35 mm in diameter (Jaffe and Horwitz, 1992).

Cells not induced to form primary lens fibres, form the cuboidal monolayer of lens epithelial cells. Only cells lying within a narrowly defined area positioned anteriorly to the lens equator, known as the germinative or proliferation zone (figure 1.2, B), are able to divide and differentiate (Jaffe and Horwitz, 1992). After the 25 mm stage (week 8) of development, secondary lens fibres are formed through division of cells within the germinative zone and are continued to be produced throughout life (Brown and Bron, 1996). During the elongation process, lens cells expel their nuclei, mitochondria, golgi bodies, and rough and smooth endoplasmic reticulum leading to mature lens fibres that are unable to divide and have minimal turnover of protein constituents (McAvoy, 1981).

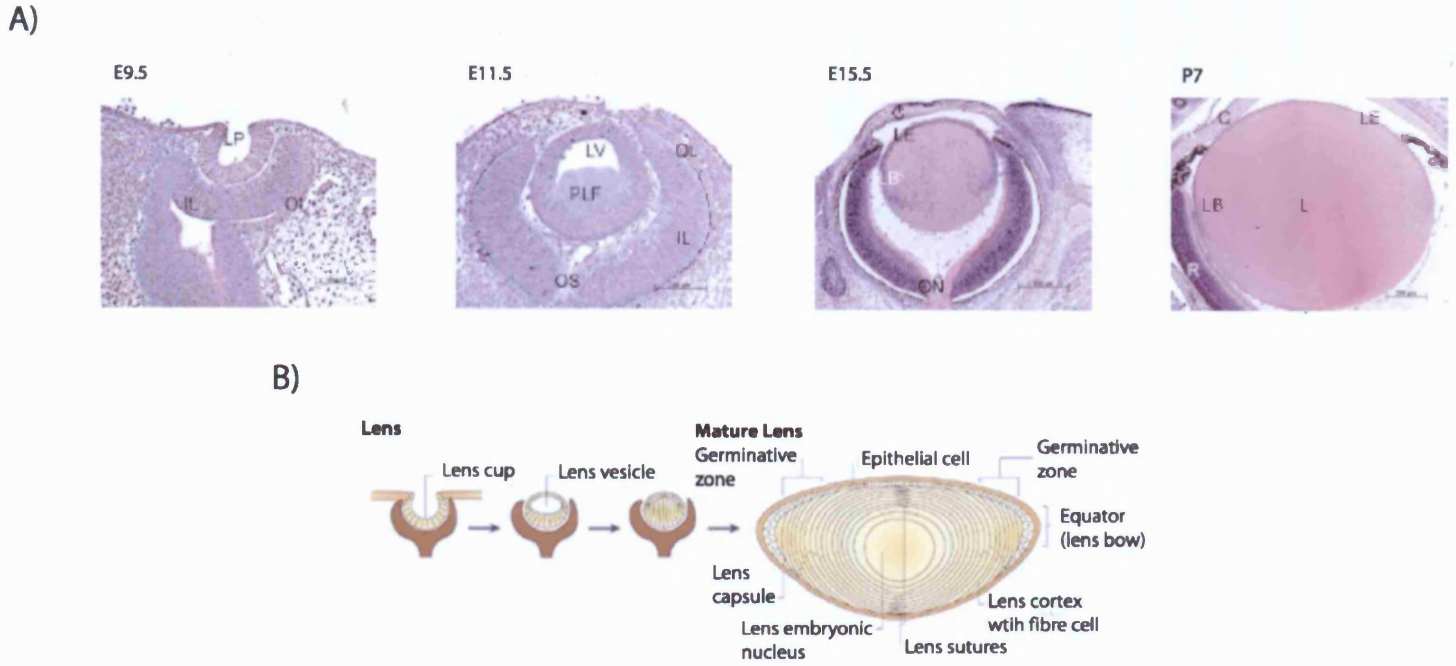


Figure 1.2 The developing lens

A) A developing mouse eye lens shown as histological sections at embryonic day (E) 9.5, E11.5, E15.5, and Postnatal day (P) 7. At E9.5 the lens placode (formed from the thickening of the surface ectoderm, overlying the optic veiscle) invaginates forming the lens pit (LP). At E11.5 the lens pit has closed over to form the lens vesicle (LV). The lens vesicle detaches from the surface ectoderm and the primary lens fibres (PLF) begin to elongate from the posterior part of the lens, filling the lumen. By E15.5, the PLF have completely filled the lumen, which is followed by division and migration of epithelial cells to the lens equator (LE). At the lens bow (LB) they differentiate into secondary lens fibres cells and elongate to form lamellae surrounding the embryonic nucleus. At P7 the eye is developed and the lens (L) is free of cellular organelles. IL, inner layer; OL, outer layer; OS, optic stalk; C, cornea; NFL, nerve fibre layer; ON, optic nerve; R, retina. B) Following the PLF completely filling the lumen of the lens vesicle, secondary lens fibres (SLF) elongate at the lens bow region. The points at which the SLF meet at the posterior and anterior poles gives rise to the lens sutures. Taken from Graw and Loster, 2003.

Sutures are formed at points where secondary lens fibres come into apposition (figure 1.2; Kuszak *et al.*, 1984). Sutures start to appear at about the 35 mm stage (8-9 weeks of fetal life). Until birth, addition of lens cell layers gives rise to a suture plane, with a symmetrical Y pattern in the anterior section and a symmetrical inverted Y pattern at the posterior section. Following birth, the formation of sutures becomes complex and irregular leading to various zones of discontinuity (Jaffe and Horwitz, 1992).

The human lens capsule is a transparent basement membrane, completely encompassing the lens, and continues to grow throughout life. At the 13 mm developmental stage (start of week 7), epithelial cells secrete the anterior portion of the capsule whilst the elongating fibre cells secrete the posterior portion (Jaffe and Horwitz, 1992). The lens capsule plays an important role in modelling the shape of the lens as its volume continues to increase throughout the life of the lens.

1.3 Morphology of congenital cataracts

Cataract is an opacity of the lens, which varies in morphology, can be either static or progressive, and is frequently confined to a specific area of the lens. Inherited congenital cataract can occur as an isolated anomaly or may be associated with an inherited disease syndrome. Examples of X-linked inherited syndromic cataract and their loci include, Alport syndrome at Xq22 (Atkin *et al.*, 1988; Szpiro-Tapia *et al.*, 1988), Lowe Syndrome at Xq26.2 (Silver *et al.*, 1987; Nussbaum *et al.*, 1997; Reilly *et al.*, 1990), Oculofaciocardiodental syndrome at Xp11.4, and Nance-Horan Syndrome at Xp22.2 (Lewis, 1989; Stambolian *et al.*, 1990; Zhu *et al.*, 1990; Wilkie *et al.*, 1993; Bergen *et al.*, 1994; Aalfs *et al.*, 1996; Toutain *et al.*, 1997).

Variation of type and severity of lens opacities is great. They range from dense, total opacities involving different structures of the lens leading to total blindness, to a minute white dot in the anterior capsule (Amaya *et al.*, 2003). Only those of relevance to this project are referred to below.

1.3.1 Total cataracts

These types of cataract have been reported in families with autosomal dominant and X-linked modes of inheritance (Semina *et al.*, 1998; Francis *et al.*, 2002).

Cataracts affecting the whole of the lens usually have an early onset, requiring early surgical intervention if they have a dramatic effect on vision. General opacity of all lens fibres is commonly referred to as a total cataract. Total cataract may develop from lamellar or nuclear cataracts and in some cases the lenses are found to be totally opaque upon first diagnosis (figure 1.3). Whole lens cataracts are seen in Down syndrome and syndromes, as well as familial or sporadic cases. Less common are congenital Morgagnian cataracts. These are total, dense cataracts in which the outer zones of the lens become liquefied whilst the nucleus remains intact. A third type of total cataract are disk-like and membranous cataracts, representing varying stages of reabsorption of the lens leaving either a disk of lens material or a bag of milky or crystalline substance that is either dense and completely opaque or thin and transparent (Amaya *et al.*, 2003).

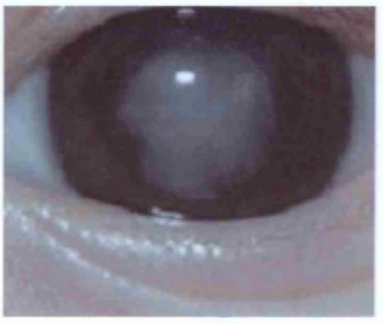


Figure 1.3 Total cataract (Taken from Reddy *et al.*, 2004)

1.3.2 Central cataracts

Nuclear cataracts consist of opacities of the majority of the embryonic or fetal nucleus (figure 1.2, B). They are static and vary from fine dots to a dense white and chalk-like, central cataract. The cataract is usually bilateral and is commonly associated with opacified cortical fibres encircling the nuclear opacity, known as riders. Nuclear cataracts are also on occasion combined with opacities of the sutures (Amaya *et al.*, 2003). Other central cataracts include lamellar or zonular cataracts, central pulverulent cataracts, cortical cataracts and cerulean, floriform or coronary cataracts.

1.3.3 Sutural cataracts

Sutural cataracts (figure 1.4) consist of opacities around or involving the sutures of the fetal nucleus (figure 1.2), are very common, stationary and usually bilateral and familial. Their morphology ranges from an increased density of the sutures to a variety of whitish or cerulan dots clustered around the sutures which may progress and form nuclear or central cataracts (Amaya *et al.*, 2003). Sutural cataracts may be the only phenotype observed in asymptomatic relatives. They are reported to be inherited in autosomal dominant and X-linked recessive modes and are seen in female carriers of Nance-Horan Syndrome (Zhu *et al.*, 1990).

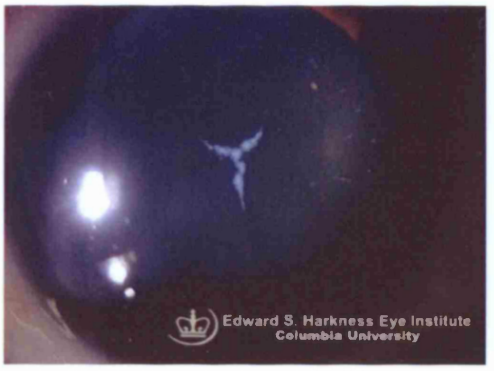


Figure 1.4 Sutural cataract (Taken from http://dro.hs.columbia.edu/antsutural.htm)

1.4 Genetics of congenital cataract

A large proportion of inherited non-syndromic cataract exhibits an autosomal-dominant mode of inheritance (table 1.1), although autosomal recessive and X-linked inheritance have been reported (Francois, 1982; Francis *et al.*, 2002). A number of chromosomal structural abnormalities have also been associated with congenital cataract such as, Wilms tumour, aniridia, genitourinary anomalies, and mental retardation (WAGR syndrome, Pinna *et al.*, 2004).

Cataract is an extremely heterogenous disease and the number of genes and mapped loci associated with cataract is ever increasing (tables 1.1, 1.2, and 1.3). Genes implicated in cataract, include those encoding transmembrane and structural proteins, and transcription factors.

1.4.1 Transmembrane proteins in the lens

Terminally differentiated lens fibre cells are metabolically inactive and therefore are dependent on the lens epithelium for adequate nutrition. The removal of waste products is also essential to the survival of the lens fibres. The presence of fully functional gap junctions and channels is imperative to prevent the precipitation of cell structural proteins, causing cataract formation (He and Li, 2000).

1.4.1.1 Connexins

In 1968, Donahue and colleagues assigned the Duffy blood group to chromosome one, such that, the cosegregation of cataract with the Duffy blood group locus became the first genetically linked autosomal disease in man (Renwick and Lawler, 1963). The Duffy Blood group locus was later refined to 1q22-23 (Matthew *et al.*, 1994) aiding the identification of a mutation in the connexin 50 gene (*GJA8*) in a family with autosomal dominant pulverulent cataract (Shiels *et al.*, 1998). *GJA8* is primarily and abundantly expressed in human lens (White *et al.*, 1992).

Connexin proteins mediate intercellular transportation of small biomolecules (<1 kDa), such as second messengers, ions and metabolites (Goodenough, 1992). In addition, connexin proteins allow adjacent cells to coordinate their activities and movements.

Consequently, their function in the lens is important for homeostasis and maintaining lens fibre transparency (Goodenough, 1992).

Connexins form a superfamily of proteins, which associate to form heterogeneous oligomeric transmembrane structures called connexons. Connexons form between lens fibre cells allowing the flow of ions, metabolites, and second messengers (Goodenough, 1980), as well as preserving lens fibre cell homeostasis (Mathias *et al.*, 1997). Of the connexins, connexin 43 (encoded by *GJA1*), connexin 46 (encoded by *GJA3*) and connexin 50 (encoded by *GJA8*) are expressed within the lens. There have been many reports of connexins (when mutated or knocked out) causing a wide spectrum of phenotypes. For example, absence of connexin 43 expression leads to dominantly inherited severe heart malformations (Reaume *et al.*, 1995), loss of connexin 50 leads to dominant zonular pulverulent cataracts (White *et al.*, 1998) and a mutation in connexin 46 was found in two families with dominant pulverulent cataract (table 1.1; Mackay *et al.*, 1999).

1.4.1.2 Aquaporin/MIP

The Major Instrinsic Protein, MIP (MIP26/AQP0) present in lens fibre cells, was the first member of the aquaporin family identified. Aquaporins facilitate the transport of water across the plasma membrane of cells by acting as channel proteins (Heymann *et al.*, 1998). Expression of *MIP* has been found to be strongest in elongating fibre cells in the bow region of the lens (Shiels and Griffin, 1993). Mutations in the *MIP* gene have been identified in two unrelated families (Berry *et al.*, 2000, table 1.1). Each family had a different missense mutation giving rise to two distinct forms of cataract. One family exhibited polymophic bilateral discrete progressive punctate lens opacities limited to mid and peripheral lamellae with additional asymmetric polar opacification. The second family were reported to have non-progressive lamellar cataract (Berry *et al.*, 2000). Functional studies implied that the mutated proteins were not targeted to the cell membrane, which may have been a result of misfolding.

Mode of Inheritance	Gene	Nomenclature	Locus	Phenotype	OMIM	Reference
Autosomal Dominant	*Beaded filament structural protein 2, Phakinin*	*BFSP2*	3q22.1	Congenital and Juvenile Y-sutural	MIM:604219	Conley *et al.*, 2000 Jakobs *et al.*, 2000 Zhang *et al.*, 2004
	Crystallin, beta A1	*CRYBA1*	17q11.2	Congenital zonular, with sutural opacities	MIM:600881	Kannabiran *et al.*, 1998
	Crystallin, beta B2	*CRYBB2*	22q11.23	Congenital cerulean type 2	MIM:601547	Litt *et al.*, 1997
	Crystallin, beta B1	*CRYBB1*	22q12.1	Pulverulent cataract	MIM:600929	Mackay *et al.*, 2002
	Crystallin, alpha A	*CRYAA*	21q22.3	Congenital Zonular central nuclear	MIM:123580, 604219	Litt *et al.*, 1998
	Crystallin, alpha B	*CRYAB*	11q23.1	Congenital posterior polar	-	Berry *et al.*, 2001
	Crystallin, gamma C	*CRYGC*	2q33.3	Coppock-like/ variable zonular	MIM:123680	Heon *et al.*, 1999
	Crystallin, gamma D	*CRYGD*	2q33.3	Punctate, progressive juvenile-onset	MIM:123690	Stephan *et al.*, 1999 Heon *et al.*, 1999
				Congenital crystalline aculeiform cataract	MIM:115700	
	Crystallin, gamma S	*CRYGS*	3q27.3	Progressive cortical cataract	-	Sun *et al.*, 2005
	Gap junction protein alpha 3, 46 kD (Connexin 46)	*GJA3*	13q12.11	Zonular pulverulent type 3	MIM:601885	Mackay *et al.*, 1999
	Gap junction protein alpha 8, 50 kD (Connexin 50)	*GJA8*	1q21.1	Zonular pulverulent type 1	MIM:116200	Shiels *et al.*, 1998
	Major Intrinsic protein	*MIP*	12q13.3	Polymorphic/Lamellar	MIM:154050	Berry *et al.*, 2000
	Heat shock transcription factor 4	*HSFA*	16q21	Lamellar cataract	MIM:116800	Bu *et al.*, 2002
	v-Maf musculoaponeurotic fibrosarcoma oncogene homologue	*MAF*	16q23.2	Juvenile –onset +/- microcornea/coloboma	MIM:177075	Jamieson *et al.*, 2002
	Paired-like homeodomain transcription factor 3	*PITX3*	10q24.32	Total	MIM:602669	Semina *et al.*, 1998
	Paired-like homeobox-containing gene 6	*PAX6*	11p13	Congenital cataract	-	Glaser *et al.*, 1994

Table 1.1 Disease genes implicated in autosomal dominant cataract

Mode of Inheritance	Gene	Nomenclature	Locus	Phenotype	OMIM	Reference
Autosomal Recessive	*Lens Intrinsic membrane protein 2 (19 kD)*	*LIM2*	19q13.41	Late-onset pulverulent cortical	MIM:154045	Pras *et al.*, 2002
	Heat shock transcription factor 4	*HSFA*	16q21	Congenital Congenital, total white cataract	-	Forshew *et al.*, 2005, Smaoui *et al.*, 2004
	Crystallin, alpha A	*CRYAA*	21q22.3	Congenital	-	Pras *et al.*, 2000
	Crystallin, beta B3	*CRYBB3*	2q11.23	Congenital, nuclear with cortical riders	-	Riazuddin *et al.*, 2005a
	Glucosaminyl (N-acetyl) transferase 2	*GCNT2*	6p24	Congenital	-	Pras *et al.*, 2004
Syndromes	*Ferritin Light Chain*	*FTL*	19q13.33	Hyperferritinemia-cataract syndrome	MIM:600886	Beaumont *et al.*, 1995
	Oculocerebrorenal syndrome of Lowe	*OCRL*	Xq25	Congenital	MIM:309000	Leahey *et al.*, 1993
	Protein-O-mannosyltransferase 2	*POMT2*	14q24.3	Lamellar	-	Van Reeuwijk *et al.*, 2005
	BCL-6 interacting corepressor	*BCOR*	Xp11.4	Congenital cataract	MIM:300166	Ng *et al.*, 2004

Table 1.2 Disease genes implicated in autosomal recessive cataract and syndromal cataract.

These mutants were also found to have a dominant-negative effect on wildtype protein (Francis *et al.*, 2000). The Cat^{Lop} mouse mutant (Lens opacity) also has a single amino acid substitution in the *Mip* gene, which inhibits targeting of its protein product Mip, to the cell membrane (Shiels and Griffin, 1993; Shiels and Bassnett, 1996).

1.4.1.3 Lens intrinsic membrane protein 2, Lim2

The lens intrinsic membrane protein 2 (Lim2, also referred to as MIP19) is one of the most abundant integral membrane protein in lens fibre cells. Its role within the lens is believed to involve the exchange of ions and metabolites between lens fibres, epithelial cells, and the extracellular matrix. Expression analysis of the *Lim2* gene in mouse embryos identified its presence in the head region from embryonic day 12 leading to the hypothesis that it plays an important role in lens development (Zhou *et al.*, 2002). A missense mutation in the human *LIM2* gene was identified in a family with autosomal recessive presenile cataract (Pras *et al.*, 2002; table 1.2). A missense mutation in the mouse *Lim2* gene, results in a total opacity with a dense cataract, which is also associated with micophthalmia in homozygotes (Steele *et al.*, 1997).

1.4.2 Structural proteins of the lens

1.4.2.1 Crystallins

Crystallins are contained within lens fibre cells. They maintain lens fibre transparency through the formation of tightly packed stable oligomers, which interact with the surrounding cytoskeleton (Carter *et al.*, 1995). The three major classes of crystallins (α-, β-, γ-) account for the majority of proteins of the human lens with approximately 90% of water-soluble proteins being crystallins. The β- and γ- crystallin polypeptides are part of a superfamily of proteins. They share a common two-domain structure, each domain consisting of four torqued β-pleated sheets referred to as "Greek key" motifs (Lubsen *et al.*, 1988). The α- crystallin family includes αA- and α-B crystallin and both are considered to be molecular chaperones and members of the small heat-shock protein family (Graw, 1997).

Numerous mutations in the crystallin genes have been implicated in autosomal dominant cataract (table 1.1), the most recent being the *gamma-S crystallin* gene (*CRYGS*). A missense mutation in exon 2 of the *CRYGS* gene was found to co-segregate with dominant progressive cortical cataracts in a large six-generation family (Sun *et al.*, 2005). Another crystallin gene recently implicated in autosomal recessive cataract was crystallin beta, B3 (table 1.2; Riazuddin *et al.*, 2005a).

The α-crystallin complexes are composed of two subunits: αA- and αB-crystallin, encoded by two genes *CRYAA* and *CRYAB*, respectively. Both αA- and αB-crystallin are highly expressed in the developing and mature lens. Mutations in the *CRYAA* gene have been implicated in both autosomal dominant and autosomal recessive forms of cataract (tables 1.1 and 1.2; Litt *et al.*, 1998; Pras *et al.*, 2000).
A mutation in the *CRYAB* gene has been reported to cause autosomal dominant congenital posterior polar cataract (Berry *et al.*, 2001).

The β- and γ- crystallins are coded for by a β/γ crystallin superfamily consisting of six β- and seven γ-crystallin genes (*CRYB* and *CRYG*, respectively). The β-crystallins constitute the majority of the water soluble protein in the lens. In mammals the γ-crystallins are expressed in the lens and retina (Jones *et al.*, 1999). A number of mutations within the β-crystallin genes are implicated in autosomal dominant cataract (table 1.1). Mutation of the *CRYBB2* gene has been found to cause congenital cerulean cataract (Litt *et al.*, 1997), whereas mutation of the CRYBB1 is associated with pulverulent cataract (Mackay *et al.*, 2002). A splice site mutation in *CRYBA1*, which codes for βA3 crystallin, has been found to cause dominant zonular cataract with sutural opacities (Kannabiran *et al.*, 1998). A missense mutation in the *CRYGD* has been associated with autosomal dominant "coral-like" cataract (Mackay *et al.*, 2004).

1.4.2.2 Cytoskeletal proteins

Crystallin proteins interact with the cytoskeleton of lens cells thereby maintaining their structural framework. Some of the cytoskeletal proteins found within the lens include: Filensin, vimentin, and CP49 (phakinin/beaded filament structural protein 2). The two intermediate filament proteins CP49 and filensin form the beaded filament, a structure unique to the lens. Beaded filaments were first identified back in 1972 by Maisel and

Perry, and have since been found to be the major cytoskeletal component of the lens. Filensin and CP49 are encoded by the *BFSP1* and *BFSP2* genes, respectively. Mutations within the *BFSP2* gene have been associated with autosomal dominant cataracts in humans (table 1.1; Conley *et al.*, 2000; Jackobs *et al.*, 2000). The cataract phenotype observed from mutations in the *BFSP2* gene, were of a juvenile-onset progressive cataract in one family (Conley *et al.*, 2000). Apart from two affected individuals exhibiting congenital nuclear cataract, the majority of the affected members from this family had lamellar cataract. The second family with mutations in the BFSP2 gene had pulverulent and spoke-like cortical opacities from birth (Jakobs *et al.*, 2000). Recently, Zhang and colleagues reported that a mutation within the BFSP2 was the cause of Y sutural cataract in a Chinese family (table 1.1; Zhang *et al.*, 2004).

1.4.2.2.1 Oculocerebrorenal Syndrome of Lowe (OCRL)

Lowe Oculocerebrorenal Syndrome (OCRL, MIM 309000) is a rare X-linked disorder characterised by bilateral congenital cataract, neonatal hypotonia, severe mental retardation, and renal Fanconi syndrome (a defect in solute and protein re-absorption in the proximal tube, resulting in progressive kidney wasting and eventually end-stage renal disease). Mutations within the *OCRL1* gene have been identified in patients with Lowe Oculocerebrorenal syndrome (Suchy *et al.*, 1995; Zhang *et al.*, 1995). The *OCRL1* gene encodes a phosphoinositide 5-phosphatase, with preference for the substrate phosphatidylinositol 4,5-bisphosphate $PI(4,5)P_2$ and Phosphatidylinositol 3,4-5-trisphosphate $PI(3,4,5)P_3$ (Zhang *et al.*, 1995; Schmid *et al.*, 2004).

Two protein isoforms, A (901 aa) and B (893 aa), which differ by 8 amino acids (EDSFLEKE), are coded for by *OCRL1*. Both isoforms are expressed in all tissues with the exception of brain, in which only isoform A is expressed (Johnson *et al.*, 2003). Within the OCRL1 protein sequence are two conserved domains; a Rho GAP domain located at the C-terminus and an inositol polyphosphate 5-phosphate domain (figure 1.5).

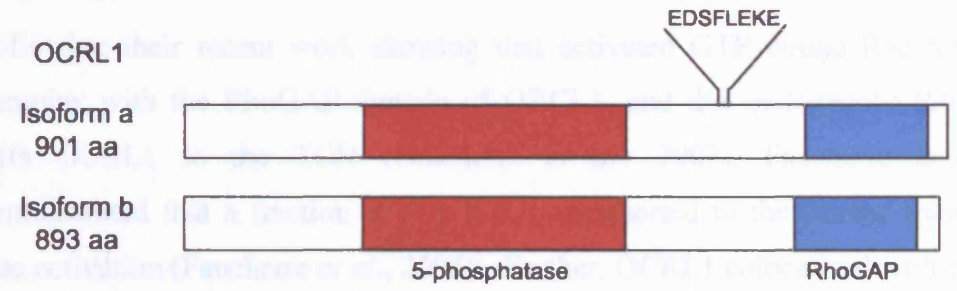


Figure 1.5 The human Oculocerebrorenal syndrome of Lowe 1 (OCRL1) protein

The *OCRL1* gene encodes two isoforms, A (901 aa) and B (890 aa), which differ by the presence/absence of eight amino acids (EDSFLEKE). The 5-phosphatase and RhoGAP domains are highlighted. Adapted from Lowe, 2005.

Kidney proximal tubules and fibroblast cell lines derived from Lowe Oculocerebrorenal syndrome patients and therefore devoid of OCRL1, have an elevated concentration of phosphatidylinositol 4,5-bisphosphate PI(4,5)P_2 (Zhang *et al.*, 1998; Wenk *et al.*, 2003). PI(4,5)P2 has a crucial role in cell signaling, polymersiation of the actin cytoskeleton, and protein trafficking (Berridge and Irvine, 1989; Toker, 1998; Raucher *et al.*, 2000).

The first observation that the actin cytoskelton of fibroblasts was abnormal in Lowe Oculocerebrorenal syndrome patients, was made by Suchy and Nussbaum in 2002. Suchy and Nussbaum described the actin cytoskeleton in these fibroblasts to exhibit a reduced number and length in actin stress fibres, along with an altered response to depolymerising agents and an altered distribution of gelsolin and α-actinin, two actin-binding proteins.

Over the last ten years, the OCRL1 protein has been localised to the *Trans*-Golgi network (TGN) of fibroblasts and epithelial cells (Suchy *et al.*, 1995; Dressman *et al.*, 2000) and more recently, OCRL1 has been implicated in regulating protein trafficking between the TGN and endosomes (Choudhury *et al.*, 2005). OCRL1 was found to be present in clathrin-coated vesicles operating between the TGN and endosomes, and has been shown to bind the terminal domain of clathrin-heavy chain (CHC) through its

LIDLE motif residing between the 5-phosphatase and RhoGAP domains (ter Haar *et al.*, 2000; Ungewickell *et al.*, 2004; Choudhury *et al.*, 2005).

Following their recent work showing that activated GTP bound Rac forms a stable complex with the RhoGAP domain of ORCL1, and that endogenous Rac colocalizes with OCRL1 to the TGN (Faucherre *et al.*, 2003), Faucherre and colleagues demonstrated that a fraction of ORCL1 is transported to the plasma membrane upon Rac activation (Faucherre *et al.*, 2005). Further, OCRL1 colocalized with polymersized actin in membrane ruffles suggesting a role in regulating actin polymersiation (Faucherre *et al.*, 2005). $PI(4,5)P_2$ was also found to accumulate in plasma membrane ruffles of patient dermal fibroblasts compared with controls (Faucherre *et al.*, 2005). Whether the RhoGAP domain of OCRL1 is a fully functional GTPase-activating domain is under question. Firstly, the RhoGAP domain of OCRL1 has a glutamine in place of the critical catalytic arginine residue in other Rho GAPs. Secondly, binding of the RhoGAP domain in OCRL1 to GDP- and GTP-bound Rac was found to be of equal, with only weak GAP activity to the GTP-bound form (Jefferson and Majerus, 1995).

1.4.3 Transcription factors

Transcription factors regulate the formation of the lens throughout development.

1.4.3.1 Pituitary homeobox 3 (PITX3)

Mutation of the *pituitary homeobox 3* (*PITX3*) gene, a member of the RIEG/PITX homeobox gene family, has been found to cause anterior segment mesenchymal dysgenesis (ASMD) and congenital cataracts in humans (table 1.1; Semina *et al.*, 1998). A 17-bp insertion in the 3' –end of the coding sequence and resulting in a frame shift, was identified in a patient with ASMD and cataract. In addition a G→A transition in exon 2 (resulting is a serine being replaced for asparagine) was identified in a patient with congenital cataract.

1.4.3.2 Paired-like homeobox-containing gene 6 (PAX6)

Paired-like homeobox-containing gene 6 (*PAX6*) is essential for early eye development, specification of ocular tissues and normal eye development in vertebrates (Hill *et al.*, 1992). PAX6 contains a paired domain/homeodomain and is believed to have a role in regulating lens crystallin expression (Duncan *et al.*, 1998). Mutations in the human *PAX6* gene have been found to cause congenital cataracts, aniridia, anophthalmia and central nervous system defects (table 1.1; Glaser *et al.*, 1994; Jordan *et al.*, 1992; Hanson *et al.*, 1994; Martha *et al.*, 1995).

1.4.3.3 The bZIP transcription factor family (MAF)

The MAF family are basic leucine zipper (bZIP) transcription factors expressed in the lens placode and primary lens fibres (Shimada *et al.*, 2003). MAF proteins bind as homo- or heterodimers to MAF responsive elements (MAREs) found in the promoters of the crystallin encoding genes and *PITX3* (Ogino and Yasuda, 1998; Semina *et al.*, 2000).

A family reported to have cataract associated with microcornea and iris coloboma (table 1.1), were found to harbour a mutation affecting the DNA-binding domain of the MAF gene (Jamieson *et al.*, 2002). The cataract phenotype within this family was one of pulverulent, autosomal juvenile-onset (Jamieson *et al.*, 2002).

1.4.3.4 Heat shock transcription factor 4 (HSF4)

A regulator of small and large heat shock proteins is the heat shock transcription factor 4 (HSF4). Small heat shock proteins such as αB crystallin have an inherent chaperone activity that is not ATP-dependent, unlike the larger heat shock proteins. Larger heat shock proteins also have the capability of refolding denatured proteins and protecting those newly synthesized proteins from misfolding (Bagchi *et al.*, 2002). Mutation of the DNA binding domain of HSF4 is associated with autosomal dominant lamellar cataract (table 1.1; Bu *et al.*, 2002). More recently, mutation of *HSF4* has been implicated in autosomal recessive congenital cataract (table 1.2; Smaoui *et al.*, 2004; Forshew *et al.*, 2005).

1.4.3.5 BCL-6 interacting corepressor (BCOR)

Recently, several families with Oculofaciocardiodental (OFCD, OMIM 300166) were found to contain a range of mutations (frameshifts, deletions, and nonsense mutations) in the *BCL-6 interacting corepressor* (*BCOR*) gene (Ng *et al.*, 2004). OFCD is a distinct form of syndromic microphthalmia associated with congenital cataracts, narrow face, broad nasal tip with separated cartilage, cleft palate, and cardiac and dental anomalies. OFCD is inherited in an X-linked dominant pattern with male lethality. In addition, a family with Lenz micropthalmia (*MAA2*) was found to have a missense mutation in the *BCOR* gene, substituting a conserved proline (Ng *et al.*, 2004). The authors proposed that the two diseases result from defects in alternative functions of the BCOR protein (Ng *et al.*, 2004).

1.4.4 Cataract disease loci

In addition to the aforementioned genes already implicated in autosomal dominant, recessive and syndromic forms of cataract, a number of disease loci have been mapped for which a novel disease gene is yet to be identified. Table 1.3 summarises cataract disease loci mapped over the last few years.

Mode of Inheritance	Locus	Phenotype	Reference
Autosomal dominant	1p36.13-36.21	Total congenital cataract	McKay *et al.*, 2005
	2p35-pter	Coralliform cataract	Gao *et al.*, 2005
	15q22.33-24.2	Anterior polar cataract (associated with keratoconus)	Hughes *et al.*, 2003
Autosomal Recessive	9q13-22	Congenital cataract	Forshew *et al.*, 2005
	19q13	Nuclear cataract	Riazuddin *et al.*, 2005b

Table 1.3 Recently mapped disease loci for autosomal dominant and recessive forms of cataract

1.5 The Human and Mouse Genome Projects

1.5.1 The Human Genome Project and the role of bioinformatics

A major achievement of the human genome project was the publication of the draft sequence in the February 2001 issue of Nature (The Human Genome, 2001, Nature 409: 745-964). Subsequently, international efforts concentrated on converting the draft sequence into finished sequence, which was announced by the human genome sequencing consortium in April 2003. The overall goal of the Human Genome project was to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of the genes and genetic markers (Wolfsberg *et al.*, 2001). The International Human Genome Sequencing Consortium constructed the working draft by putting together sequence segments derived from over 20,000 large-insert clones (YACs, BACs, and PACs), and these results were made available on the web by numerous international centres including the Sanger Centre (www.sanger.ac.uk), the University of California at Santa Cruz (UCSC; www.genome.ucsc.edu), and The National Centre for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov/genome/guide/human/).

The finished sequence covers 99 % of the human genomes gene-containing regions (euchromatin) sequenced to an accuracy of 99.99 %. Finished sequence is a technical term meaning that the sequence is highly accurate (less than 1 error per 10,000 letters), and highly contiguous (with remaining gaps corresponding to difficult regions including 400 stretches of repetitive DNA and centromeres). The average DNA letter now sits on a stretch of 27,332,000 base pairs of uninterrupted, high quality sequence (~334 × longer than the working draft; Pennisi, 2003).

Closing the gaps in the human genome will not be completed through the industrial-scale efforts of the Human Genome Project, but instead will be slowly filled through individual research projects and the development of new technologies. Interestingly, the total number of genes within the human genome is yet to be determined, with rough estimates of 30,000 genes compared to the 35,000 – 45,000 estimates from the release of the draft sequence (Pennisi, 2003), and pre-draft estimates of about 100,000.

A number of resources are available for analysing the human genome. There are currently at least three sites, NCBI, Ensembl or UCSC, displaying genomic viewers and web-searchable datasets which allow analysis of the human genome sequence without the need to run complex software locally (Birney *et al.*, 2001, Kent *et al.*, 2002). The sequence is continually being functionally annotated by the human genetics community, in search of molecular mechanisms of genetic diseases by means of gene prediction programs, SNP detection (for analysis of human genetic diversity and to identify genes for polygenic diseases), and characterisation of novel transcripts. More sophisticated software programs have been developed to fully utilise the genome sequence as it becomes more contiguous. BLAT, a DNA/Protein Sequence Analysis program, written by Jim Kent at UCSC identifies the location of a query sequence within a genome. BLAT is designed to quickly find sequences of 95% and greater similarity and is especially useful for establishing exon/intron organisation of cDNA sequences using the latest genome freeze available.

A wealth of data and software is available that can be used to predict genes. Information on expressed sequences (i.e. expressed sequence tags (ESTs) defined from complementary DNAs, cDNAs) and proteins from humans and other organisms can provide a fast, accurate resource for resolving gene structures against the vast genomic background; these, in conjunction with sequence alignment programs such as BLAST (Basic Local Alignment Search Tool), can determine the complete structure of a single gene (Altschul *et al.*, 1990, Birney *et al.*, 2001). With BLAST, users input either a nucleotide or amino acid query sequence, and search a nucleotide or amino acid sequence database. The program returns a list of the sequence "hits", alignments to the query sequence and statistical values (Wolfsberg and Madden, 1999).

Genes can be found through an implied relationship to something else – for example, being a putative ortholog or paralog (see section 1.8; Wolfsberg *et al.*, 2001).

1.5.2 Annotating the human genome sequence

Following the completion of a working draft for the human genome, it was imperative that the human genome be annotated fully in order to utilise the vast amount of sequence data generated. Annotation of the human genome sequence involves labelling those areas of the genome with biological significance. To annotate the human genome sequence (or other genomes), involves correctly positioning known genes on the human genome sequence, and predicting new genes. To correctly position known genes, messenger RNAs (mRNAs) are aligned with the genomic template using computer software. In addition, single nucleotide polymorphisms (SNPs) and sequence tagged sites (STSs) are also positioned within the human genome sequence.

Predicting new genes is achieved through the alignment of expressed sequence tags (ESTs), which also highlight whether a gene is alternatively spliced. The validity of a gene prediction by gene prediction software is greatly enhanced from alignment of ESTs to such novel gene predictions. Comparative genomics is also a powerful tool for identifying new genes. Confidence in gene predictions arises upon identifying an ortholog in closely related species, such as the mouse. Further, comparison of orthologs will help to resolve the genomic structures of newly identified genes.

1.5.3 The Mouse Genome Project

Since the release of the human genome draft sequence in early 2001, scientists have been keen to fully annotate the human genome. A powerful means by which this task can be undertaken involves comparative genomics. The laboratory mouse, *Mus musculus*, provides a model organism for human biology. The mouse has many similarities to humans with respect to physiological homeostasis, whole-organ systems, reproduction and more importantly disease, which are reflected in their genomes.

Human and mouse lineages diverged approximately 75-125 million years ago and as a consequence of evolution, their genome sequences have diverged at a rate of nearly one substitution for every two nucleotides through the deletion and insertion of sequences. Orthologous sequences within the two genomes can be aligned due to the slow rate at which the sequences have diverged. However, the divergence rate has been great

enough to allow functionally important elements within the two genomes to be recognised through their greater degree of conservation (Waterston *et al.*, 2002).

1.5.3.1 The Mouse Genome Sequencing Consortium (MGSC)

The MGSC, originally comprised of three large sequencing centres (the Whitehead/Massachusetts Institute of Technology (MIT) Centre for Genome Research, the Washington University Genome Sequencing Centre and the Wellcome Trust Sanger Institute), who made a concerted effort to sequence the entire mouse genome (http://www.sanger.ac.uk/Projects/M_musculus/).

The mouse genome draft sequence was generated by assembling about sevenfold sequence coverage from female mice of the C57BL/6J strain. The C57BL/6J strain was selected based on its predominant role in mouse genetics and sequence from a female mouse was selected in order to obtain equal coverage of the X chromosome and the autosomes. Consequently the Y chromosome was omitted from the genome sequencing effort.

1.5.3.2 Sequencing the mouse genome using a hybrid strategy

A hybrid strategy consisting of four components, was employed for sequencing the mouse genome:

(1) Production of a BAC-based physical map of the mouse genome by fingerprinting and sequencing the ends of clones of a BAC library
(2) Whole genome shotgun (WGS) sequencing to approximately sevenfold coverage and assembly to generate an initial draft genome sequence
(3) Hierarchical shotgun sequencing of BAC clones covering the mouse genome combined with the WGS data to create a hybrid WGS-BAC assembly
(4) Production of a finished sequence by using the BAC clones as a template for directed sequencing.

Hierarchical shotgun sequencing was employed to overcome the difficulties encountered when sequencing genomes containing highly repetitive sequences.

1.5.3.3 Conservation of synteny between human and mouse genomes

Initial comparative analysis indicates that the mouse genome is about 14 % smaller than the human genome (2.5 Gb compared to 2.9 Gb, respectively). Further, over 90 % of both the mouse and human genomes can be partitioned into corresponding regions of synteny (Waterston *et al.*, 2002).

The extent of synteny varies between the individual chromosomes except for the X chromosome, which is represented as single reciprocal syntenic blocks (Figure 1.6). Comparisons between the newly generated sequence-based map of conserved synteny and the most recent map based on 3,600 loci, has identified many more conserved syntenic segments (342 compared with 202) and a few conserved syntenic blocks (217 compared with 170). The new sequence-based map of conserved synteny highlights numerous rearrangements of segments within blocks, especially on the X chromosome.

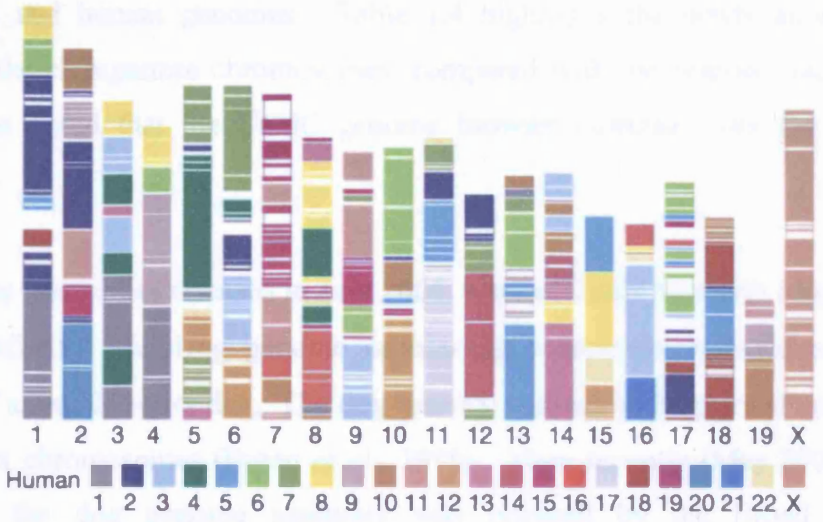


Figure 1.6 Synteny between the human and mouse genomes.

Segments and blocks > 300 Kb in size with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome. The 343 segments are separated from each other by thin, white lines within the 217 blocks of consistent colour (taken from Waterston *et al.*, 2002).

1.5.4 Other genome projects

A number of other genome projects are available for viewing and comparative analysis with the human genome. The closest human living evolutionary relative, the chimpanzee (Pan *troglodytes*), has a genome that differs by roughly 1.4 % to that of humans. The Broad Institute and Genome Sequencing Centre at the Washington University school of Medicine made the Pan *troglodytes* genome available on the public domain back in December 2003. The chimpanzee genome consists of 23 pairs of autosomes and a pair of sex chromosomes. The chimpanzee genome has been humanised in the sense that it was assembled using the human genome as a guide. Consequently, sequences such as insertions/deletions, and gene duplications are subject to incorrect representation by the current assembly (www.ncbi.nlm.nih.gov/Genomes/). A new chimpanzee chromosome naming system was adopted following a recent proposal (McConkey, 2004). This new system names the chimpanzee chromosomes to correspond to their syntenic human chromosomes, thus enabling comparisons between chimpanzee and human genomes. Table 1.4 highlights the newly adopted naming system for the chimpanzee chromosomes, compared with the original naming system. It should be noted that the UCSC genome browser currently uses the old naming system.

Two genome assemblies released in July 2004, were of Canis *familaris* (dog) and Danio *rerio* (zebrafish). The Dog genome sequencing consortium released a 7.6× WGS assembly of a female boxer dog. The dog genome consists of 38 pairs of autosomes and a pair of sex chromosomes (Breen *et al.*, 1999). More recently (May 2005), a second version of the dog genome assembly was released by the Broad Institute of MIT/Harvard and Agencourt Bioscience (www.ncbi.nlm.nih.gov/Genomes/). The Zebrafish genome project was led by the Wellcome Trust Sanger Institute, which released a 5.7 × coverage (Zv4) of the zebrafish genome. The Wellcome Trust Sanger Institute generated the zebrafish genome by whole genome shotgun and fingerprinted BAC clone sequencing (www.ncbi.nlm.nih.gov/Genomes). The zebrafish genome consists of 25 pairs of chromosomes (Endo and Ingalls, 1968; Daga *et al.*, 1996) and a genome size of 1.7 Gb (http://www.ncbi.nlm.nih.gov).

June 2005 saw the release of a 6.2 × WGS assembly of the bovine (*Bos Taurus*) genome, generated by the Human Genome Sequencing Center at Baylor College of Medicine. Genome assembly was based on the Hereford breed, and build 2.1 included the complete mitochondrial genome, derived from a Korean native cow. The Bovine genome is organised into 29 pairs of autosomes and a pair of sex chromosomes, X and Y.

Human chromosome designations	**NEW Chimpanzee chromosome designations**	**Original Chimpanzee chromosome designations**
1	1	1
2	2A	12
2	2B	13
3	3	2
4	4	3
5	5	4
6	6	5
7	7	6
8	8	7
9	9	11
10	10	8
11	11	9
12	12	10
13	13	14
14	14	15
15	15	16
16	16	18
17	17	19
18	18	17
19	19	20
20	20	21
21	21	22
22	22	23
X	X	X
Y	Y	Y

Table 1.4 Summary of the human chromosome naming system, the original chimpanzee naming system, and the new chimpanzee naming system

1.6 Identifying human disease genes through positional cloning

Identifying disease genes through positional cloning requires only the approximate chromosomal location of the gene to be known. This is in contrast with previous position-independent methods whereby knowledge of the protein product or the DNA sequence was required.

The disease must initially be genetically mapped within a family or families. Frequently, genetic mapping results in a candidate region of 10 Mb or more which is often too large for positional cloning and therefore requires refinement. For positional cloning, small disease intervals of less than 1 Mb are preferable, so that the number of candidate genes does not exceed 10 (average 10 genes per Mb). In order to refine a disease interval dense coverage of polymorphic markers across the interval is sought. Such markers (microsatellites and single nucleotide polymorphisms, SNPs) can be found through database searching for known polymorphic markers or by screening BACs, YACs and cosmids covering the candidate region. Polymorphic markers can then be typed within a family to construct individual haplotypes, which are used to identify potential crossovers between markers and disease (see section 1.6.3.1).

The release of the human genome draft sequence has had a considerable impact on disease gene identification. Prior to the availability of the human genome draft sequence, a high-resolution physical map of the disease interval had to be compiled from YACs, BACs, PACs and cosmid contigs, often a laborious task. With the advent of the human genome sequence, known genes within the candidate region can be identified through database searching and those considered as potential candidates for disease can be screened for pathogenic mutations. If no promising candidate is identified within the disease interval, novel genes are sought through a combination of computational and experimental methods (see section 1.4.1).

1.7 Genetic maps

Genetic maps are based on recombination frequencies between genetic markers at meiosis. The unit of scale for a human genetic map is the centimorgan (cM), defined as a recombination fraction (see below) of 0.01. In physical distance, 1 cM roughly corresponds to between 0.7 and 1 Mb of DNA sequence.

Unless located in the same region of the genome, genetic markers have a tendency to recombine freely at meiosis. Markers not recombining freely are referred to as linked. Genetic markers recombine during prophase of meiosis I, where pairs of homologous chromosomes synapse and exchange segments. If such an exchange (crossover) occurs between the positions of two markers (loci), two recombinant and two nonrecombinant chromatids result. Very rarely does a recombination separate loci lying in close proximity to one another on a chromosome. Conversely, the greater the distance separating two markers on a chromosome, the greater the chance of a crossover occurring leading to the separation of the two loci.

The recombination fraction (θ) is a measure of the genetic distance between two loci. If two loci segregate independently, there is a 50 % chance that the offspring will be recombinant between the two markers and a 50 % chance they will be nonrecombinant, resulting in a recombination fraction of 0.5. Loci which are syntenic and tightly linked will segregate together not resulting in recombinants and will have a recombination fraction of 0. Any two or more markers which are not tightly linked, will have a recombination fraction ranging between 0 and < 0.5.

Sets of alleles located in the same small interval of a chromosome are usually transmitted as a block (haplotype) through a pedigree because their close proximity makes it unlikely they will undergo recombination.

1.7.1 Genetic markers

Mapping human disease genes is dependent on markers. The requirement of such markers is that they are sufficiently polymorphic so that a randomly selected individual is likely to be heterozygous and therefore informative for linkage. In addition, linkage requires markers to be spaced throughout the genome at intervals $\leq$ 20 cM. The

informativness of a marker is measured by the mean heterozygosity (equation 1) of the marker in question, that is - the chance that a randomly selected person is heterozygous for that marker.

Equation 1: $H = 1 - \Sigma pi^2$

(for alleles A1, A2, A3, with gene frequencies of p1, p2, p3, the proportion of heterozygotes is equal to $1 - (p_1^2 + p_2^2 + p_3^2)$)

The types of markers currently used for mapping human disease genes are microsatellites and single nucleotide polymorphisms (SNPs). Microsatellites are tandem repeats (often $(CA)_n$ repeats) of 1 – 6 bp, consisting of 10 – 50 copies. Advantages of microsatellites over their 'minisatellite variable number tandem repeat' (VNTR) predecessors are attributed to their distribution throughout the genome (VNTRs are not evenly distributed) and the advent of the polymerase chain reaction (PCR; Saiki *et al.*, 1988).

SNPs are now the latest generation of genetic markers. Like restriction fragment length polymorphisms (RFLPs), they are a 2 allele system but do not always result in the creation or abolishment of a restriction site. They have the advantage of being scored on solid-state arrays whereas previous genetic markers have been dependent on gel electrophoresis (Wang *et al.*, 1998). SNPs are widely distributed throughout the human genome and millions of SNPs have been identified since the completion of the human genome sequence (http://www.ncbi.nlm.nih.gov/). Such a large number of SNPs compensates for their lower informativeness in comparison to microsatellites.

1.7.2 HapMap, a haplotype map of the human genome

Trying to understand the connection between heritable inter-individual variation (e.g. disease susceptibility and appearance) and variation in DNA sequence has stemmed from the completion of the human genome project. The majority of human sequence variation takes the form of SNPs and in October 2002, the International HapMap Project was launched to genotype over one million SNPs in 269 individuals from four ethnic origins (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005). The HapMap project has been divided into two phases. Phase I aimed to genotype one

or more common SNPs for every 5 Kb of genomic sequence from 269 DNA samples. Phase II aims to genotype an additional 4.6 million SNPs in each of the HapMap samples (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005). A total of 11,500 SNPs residing within the coding region of genes, and which code for different amino acids (non-synonymous cSNPs) were successfully genotyped in Phase I (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005).

Comparisons between the genome-wide resource to a more complete database of common variation, in which all common SNPs and numerous rare SNPs have been previously discovered and tested, were made (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005). Ten representative regions of 500 Kb in length were selected from the Encyclopedia of DNA Elements (ENCODE) Project (The ENCODE Project Consortium, 2004), and sequenced in 48 individuals. Identified SNPs within these 48 individuals were combined with known SNPs on the database and genotyped in all 269 DNA samples used in the HapMap project (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005).

The ENCODE data revealed that SNPs perfectly correlate with several nearby SNPs (i.e. one SNP provides information on a neighbouring SNP), and partially correlate with numerous other SNPs (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005). Thus, the presence or absence of a SNP at one site will provide information on a neighbouring SNP (serves as a proxy for other SNPs), allowing for a reduction in the total number of SNPs to be genotyped (Kruglyak, 2005, figure 1.7). Thus several hundred thousand proxies can capture the majority of common variants through strong correlations.

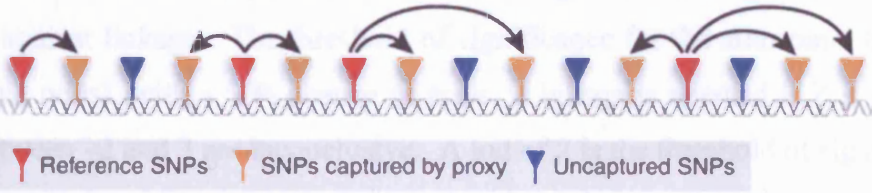


Figure 1.7 Schematic demonstrating SNPs acting as a proxy for several other nearby SNPs when testing for association with a phenotype

Red SNPs are genotyped directly, whilst the orange SNPs are captured through correlation (linkage disequilibrium) with the red SNPs. Blue SNPs are not genotyped or correlated with any genotyped SNP resulting in them not being associated with the phenotype in question. Adapted from Kruglyak, 2005.

HapMap data will allow for comprehensive genome-wide association studies to be conducted, aiding the identification of the causative genes and pathways involved in human disease. It will also help to identify alleles conferring susceptibility or resistance to common diseases, and perhaps establish inter and intra-familial phenotypic variations in disease due to differences in genetic background. This will ultimately aid in creating a better understanding of disease framework (The International HapMap Consortium, 2005, Altschuler *et al.*, 2005).

1.7.3 Linkage analysis

Identifying informative markers and typing them within families in which a mendelian disease segregates is the first step to determining whether linkage exists between those markers and the segregating disease. Seldom do families, in which a disease segregates, allow for the phase of the alleles within the heterozygous parents to be determined. Consequently, scoring each meiosis as recombinant or nonrecombinant becomes ambiguous.

Such problems are overcome by calculating the overall likelihood of linkage, on the basis that the loci are either linked (recombination fraction $\theta = 0$) or not linked (recombination fraction $\theta = 0.5$). The odds of linkage is given by the ratio of these two likelihoods, the logarithm of which is the lod score. The lod score (log of the odds, z) is the most efficient statistic for evaluating pedigrees for linkage and is calculated as a series of θ ranging from 0 to 0.5.

Linkage analysis results in a table of lod scores at various recombination fractions. Positive lod scores are evidence in favour of linkage whereas negative lod scores are evidence against linkage. The threshold of significance for the autosomal traits is $Z = 3.0$ (1000:1 odds), with a 5 % chance of error. Linkage is rejected if $Z < -2.0$ and Z values between –2 and 3 are inconclusive. A lod of 2 is the threshold of significance for establishing linkage between an X-linked trait and a marker residing on the X-chromosome. A lower threshold exists for X-linked traits due to the increased probability that two loci chosen at random should prove to be linked for only the X chromosome, compared to two loci chosen at random from the 22 autosomes.

1.7.3.1 Haplotype analysis

Once a disease gene has been mapped to a chromosomal region, it may be necessary to refine the genetic interval in order to reduce the total number of possible candidate genes. Gene rich regions may contain up to 10 genes per 1 cM of genomic DNA (approximately 1 Mb of genomic sequence). Therefore, for disease intervals greater than 1 Mb, an attempt is usually made to refine the interval through genotyping with additional informative markers within the family for which the disease interval has been mapped. Segregation of markers typed in this way, are followed through the family to try and identify any potential crossovers between the markers and the disease locus. Following the typing of additional markers, new haplotypes are constructed for the individuals within the family by minimising the total number of crossover events in each sidship.

In some circumstances there may be promising candidate genes within the mapped disease interval, which are preferentially screened and excluded before the interval is further refined by genotyping with additional markers.

1.8 Orthologs and Paralogs

The most commonplace definition for describing relationships between two entities of common decent is homology and such entities, which are homologous, are referred to as homologs. The phrase, homology, was coined by Richard Owen in 1843 when he used the phrase to designate "the same organ in different animals under every variety of form and function".

The term homolog is further divided into two subcategories, orthologs and paralogs. Genes described as orthologs are related through vertical decent (figure 1.8). They are related by speciation and are therefore derived from a single ancestral gene of the last common ancestor of the species in question. Generally, orthologs exhibit equivalent functions. Genes that are paralogs, are related through duplication (figure 1.8). Walter Fitch introduced the terms orthologs and paralogs back in 1970 (Fitch, 1970).

Events leading to gene duplication are classified into four main events:

(1) Vertical decent (speciation) with modification
(2) Gene duplication, followed by vertical decent and modification
(3) Horizontal gene transfer (HGT)
(4) Fusion, fission, and other rearrangements of genes

The first event gives rise to orthologs, and the second to paralogs (figure 1.8). The latter two events, which have been mainly recognised since the release of completed sequenced genomes, add complexity to determining whether two entities are orthologs or paralogs of one another.

Gene duplication, which gives rise to paralogs, is considered to be the main process responsible for bringing about emerging novel functions through evolution (i.e. one of the paralogs sidesteps the confines of purifying selection, thereby evolving a novel function).

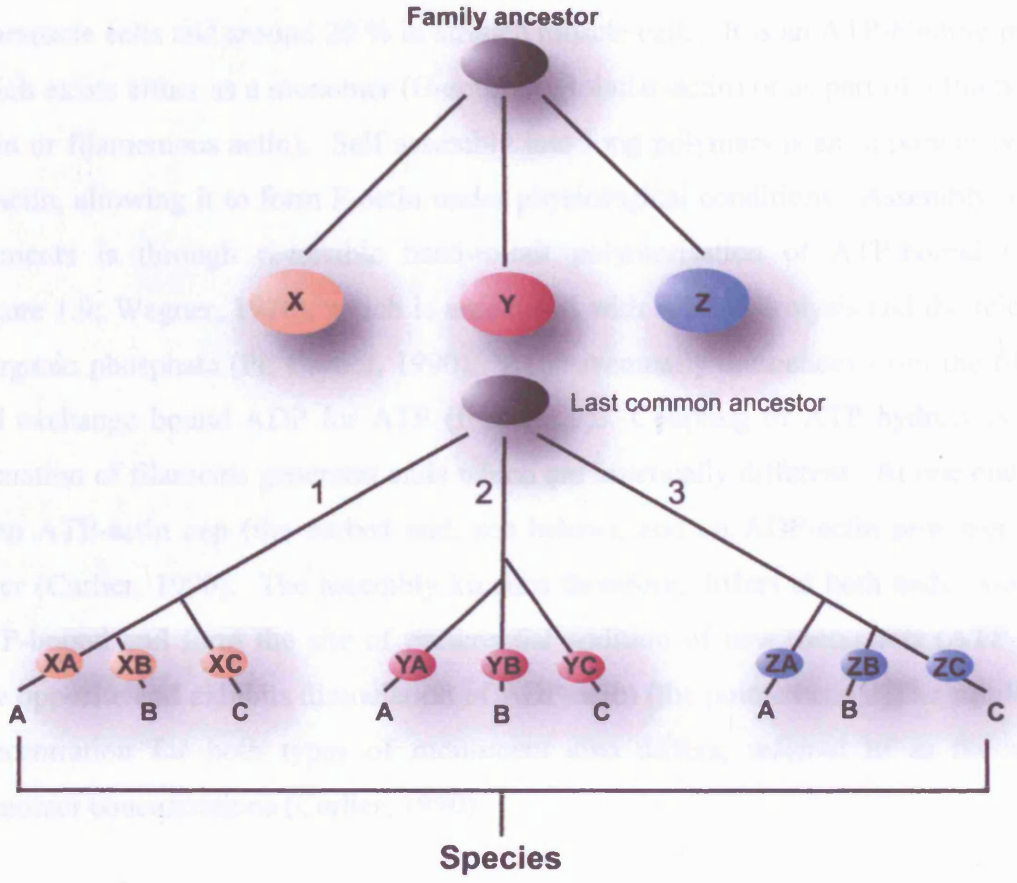


Figure 1.8 Phylogenetic tree illustrating orthologous and paralogous relationships.

A, B, and C represent different species, with each branch representing the inheritance of each gene, (X, Y, and Z) from the last common ancestor. The common ancestor of the entire gene family (genes X, Y, and Z) existed prior to the last common ancestor of all three compared species (A, B, and C). The last common ancestor of each of the three compared species already encoded three paralogous genes (X, Y, and Z). Each of these paralogs then became the progenitors of the three branches of the tree. Consequently, each gene in branch 1 is a paralog to each gene in branches 2 and 3 due to only vertical inheritance occurring via evolution from the last common ancestor. The same applies for branch 2 compared with branches 1 and 3, and branch 3 compared to branches 1 and 2. The genes in different species (A, B, and C) are orthologs to one another in each individual branch. For example, XA, XB, and XC are orthologs of each other in branch one, whilst YA, YB, and YC are orthologs to one another in branch 2.
Adapted from Koonin, 2005.

1.9 The actin cytoskeleton

The role of the actin cytoskelton within cells is diverse. Cellular events such as adhesion, division, spreading, and motility are all underpinned by the regulation of the actin cytoskeleton. Actin constitutes approximately 5-10 % of total protein in nonmuscle cells and around 20 % in striated muscle cells. It is an ATP-binding protein, which exists either as a monomer (G-actin or globular-actin) or as part of a filament (F-actin or filamentous actin). Self assembly into long polymers is an important property of actin, allowing it to form F-actin under physiological conditions. Assembly of actin filaments is through reversible head-to-tail polymerisation of ATP-bound G-actin (figure 1.9; Wegner, 1976), which is associated with ATP hydrolysis and the release of inorganic phosphate (Pi; Carlier, 1990). ADP eventually dissociates from the filament and exchange bound ADP for ATP (figure 1.9). Coupling of ATP hydrolysis to the formation of filaments generates ends which are kinetically different. At one end, there is an ATP-actin cap (the barbed end, see below), and an ADP-actin protomer at the other (Carlier, 1990). The assembly kinetics therefore, differs at both ends, where the ATP-bound end form the site of preferential addition of new monomers (ATP-actin). The opposite end exhibits dissociation of ADP-acitn (the pointed end). The equilibrium concentration for both types of monomers also differs, referred to as the critical monomer concentrations (Carlier, 1990).

Due to the nonequilibrium conditions, at high monomeric actin concentrations, molecules assemble at both the barbed and pointed ends at different rates causing the filaments to grow. The opposite is true for low monomeric actin concentrations, where the filaments shrink due to protomers dissociating from both ends. Intermediate concentrations leads to the treadmilling of actin filaments (Small *et al.*, 1993; Small, 1994; Small, 1995), and results from addition of protomers at the barbed end, and dissociation of protomers at the pointed end (figure 1.9, Neuhaus *et al.*, 1983).

Generated actin filaments are polar, containing a fast-growing end (barbed end or plus end), and a slow-growing end (pointed end or minus end). Polymersiation of actin takes place mainly at the fast-growing (barbed) end. Actin polymersiation is regulated through monomer- and filament-binding proteins, which maintain the

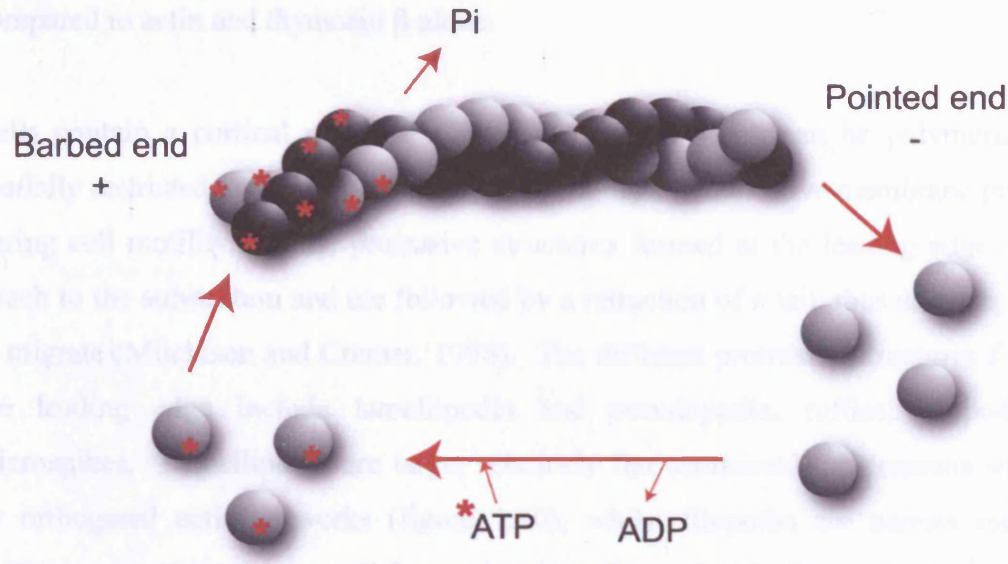


Figure 1.9 Actin polymersiation

The actin cycle shown by the incorporation of ATP-bound actin (denoted by actin monomers with a red asterisk) into the barbed end of a growing actin filament and dissociation of ADP-bound actin at the pointed end. ATP is hydrolysed by actin once incorporated into the actin filament, releasing inorganic phosphate (Pi).

monomer pool, organise filament formation and arrange filaments into arrays, and depolymerise filaments to regenerate the G-actin pool.

Members of the thymosin β family of proteins and profilins, play an essential role in maintaining a reserve of unpolymerised actin for subsequent incorporation into existing filaments or for *de novo* polymerisation. Thymosin β proteins are actin-sequestering proteins (Devineni *et al.*, 1999). When bound to actin thymosin β proteins are mobilised by the action of profilins. Profilins, which catalyse the exchange of ADP for ATP in actin monomers, sequester actin when the fast-growing end of an actin filament is capped (Goldschmidt-Clermont *et al.*, 1991). This is achieved through the binding of profilin to dissociated actin molecules from the slow growing end, keeping them in reserve. Conversely, if an actin filament is uncapped, actin-bound profilin adds to the fast-growing end as efficiently as actin alone. The affinity of profilin for actin is ten times greater than thymosin β-actin. Thus in the presence of profilin, actin is removed from thymosin β and added to the filament ends. This results in faster elongation as the

concentration of actin available for polymerisation, is higher in the presence of profilin compared to actin and thymosin β alone.

Cells contain a cortical network of actin filaments, which can be polymerised in a spatially restricted manner to generate the necessary force to drive membrane protrusion during cell motility. These protrusive structures formed at the leading edge of a cell attach to the substratum and are followed by a retraction of a tail, thus allowing the cell to migrate (Mitchison and Cramer, 1996). The different protrusive structures formed at the leading edge include lamellipodia and pseudopodia, ruffles, filopodia, and microspikes. Lamellipodia are large, relatively flat semicircular extensions supported by orthogonal actin networks (figure 1.10), whilst filopodia are narrow membrane protrusions supported by parallel actin bundles (figure 1.10). Pseudopodia are non-flat and larger in structure. Ruffles are formed from upward folding of lamellipodia, and microspikes are classed as smaller versions of filopodia. In each case, the filaments are orientated with their fast-growing (barbed) ends to the membrane (Small *et al.*, 1998). The rapid addition of actin monomers to existing filaments at the interface between the barbed end and the membrane generates the force for pushing the membrane forward. Stationary cells can often be found to form hemispherical blebs on their surface. Lamellipods and filopodia are formed from the activation of small guanosine triphosphate (GTP)-binding proteins cdc42 and Rac, respectively (Small *et al.*, 2002).

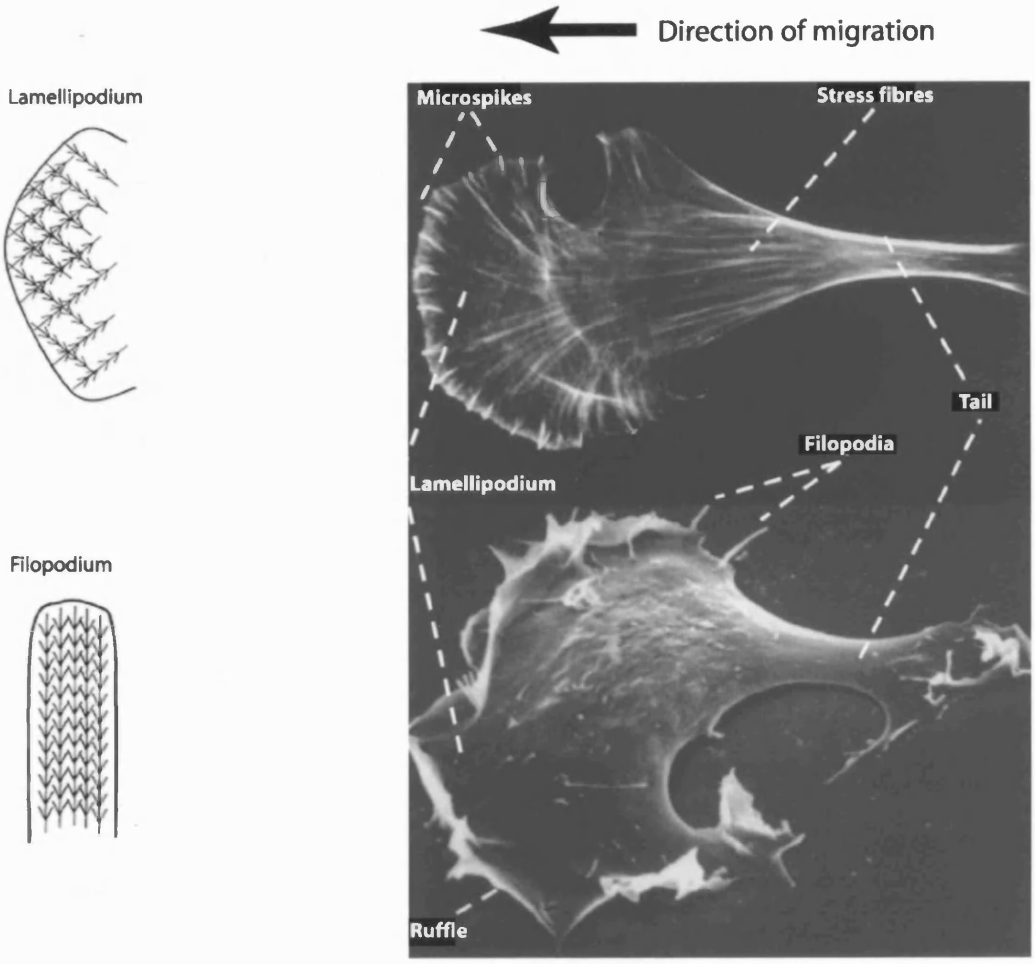


Figure 1.10 Actin within migrating fibroblasts visualised by fluorescently-labelled phalloidin (upper cell) and a scanning electron micrograph (lower cell), taken from http://www.els.net/). Stress fibres can be seen in the tails of the migrating fibroblasts. To the left of the migrating fibroblasts are schematic representations of a lamellpodium and filopodium.

1.10 The actin-related protein (Arp) 2/3 complex

The human actin-related protein (Arp) 2/3 complex is the only cellular factor known to nucleate new actin filaments with free barbed ends. Arp2/3 is composed of seven highly conserved subunits: p41, p34, p21, p20, p16, and the actin related proteins, Arp2 and Arp3 (figure 1.11, Machesky and Gould, 1999). Machesky and colleagues first isolated the Arp2/3 complex in 1994, from Acanthamoeba *castellani* (Machesky *et al.*, 1994). Since then, the Arp2/3 complex has also been isolated from a number of vertebrates including humans (Welch *et al.*, 1997), and from Saccharomyces *cerevisiae* (Winter *et al.*, 1997), and has been found to be highly conserved throughout. Although Arp2 and Arp3 are designated actin-related proteins, their actual amino acid sequence and the majority of their surface residues differ to that of actin. Arp2 and Arp3 are however, predicted to share the same fold as actin (Kelleher *et al.*, 1995).

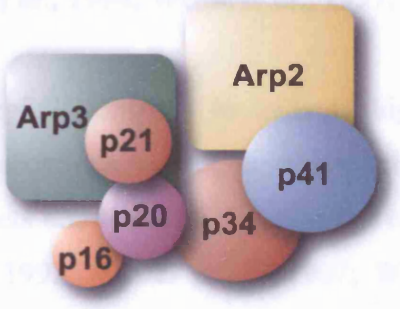


Figure 1.11 The Arp2/3 complex

The Arp2/3 complex is composed of seven highly conserved subunits: p41, p34, p21, p20, p16, and Arp2 and Arp3.

The Arp2/3 complex has an intrinsic ability to enhance actin nucleation, albeit weakly. The Arp2 and Arp3 subunits form a stable dimer that mimics actin-actin dimers, thus providing a substrate for rapid elongation of actin (Kelleher *et al.*, 1995).

The intrinsic nucleating capability of the Arp2/3 complex is enhanced through the binding of nucleation-promoting factors. Nucleation-promoting factors have been identified in the majority of eukaryotes studied to date, and in Listeria *monocytogenes* (ActA), an enteropathogenic bacterium (Welch *et al.*, 1998). The most intensively studied nucleating-promoting factors are the WASP protein family (section 1.11).

Known nucleation-promoting factors (NPFs) of the Arp2/3 complex are summarised in table 1.5.

Additional activators of the Arp2/3 complex include actin filaments themselves when in the presence of nucleation promoting factors such as the WASP protein family. Binding of the Arp2/3 complex to the side of an actin filament increases the affinity five fold for binding to a WASP protein family member (Marchand *et al.*, 2001). Such activation of the Arp2/3 complex results in the formation of Y-shaped branches on pre-existing actin filaments and is referred to as the "dendritic nucleation" model of Arp2/3 activity (Pollard *et al.*, 2000). The newly nucleated 'daughter filament' projects at a 70° angle from the original, mother filament (Mullins *et al.*, 1998; Blanchoin *et al.*, 2000). The formation of Y-shaped branches are seen in actin meshworks of leading edge lamellipodia in motile cells (figure 1.12, Svitkina and Borisy, 1999; Bailly *et al.*, 1999). Interestingly, the Arp2/3 complex is localised to the leading edge of motile cells and spreading cells (Machesky *et al.*, 1994; Welch *et al.*, 1997; Machesky *et al.*, 1997).

The importance of the Arp2/3 complex to cells is highlighted in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, in which deletion of the Arp2/3 complex leads to severe growth defects and lethality (Lees-Miller *et al.*, 1992; McCollum *et al.*, 1996; Schwob and Martin 1992; Winter *et al.*, 1997; Winter *et al.*, 1999). Further, *Drosophila Arp3*, and *Arpc1* mutants do not survive to adulthood (Hudson and Cooley 2002) and have been found to have defects in embryonic actin structures in addition to nervous system development (Zallen *et al.*, 2002).

Class of NPF	NPF	Identified in	Function	Reference
Class I	ActA	L. monocytogenes L. ivanovii	Bacterial pathogenesis, intracellular motility	Welch *et al.*, 1998
	RickA	Rickettsia conorii	Bacterial pathogenesis, intracellular motility	Gouin *et al.*, 2004
	WASP	S. cerevisiae (Las17p/Bee1p) S. pombe (Wsp1p) metazoans	Cell signalling, endocytosis	Yarar *et al.*, 1999, Winter *et al.*, 1999
	N-WASP	metazoans	Endocytosis	Rohatgi *et al.*, 1999
	Scar/WAVE	D. discoidium (Scar)	Lamellipodia formation, phagocytosis, endocytosis.	Machesky *et al.*, 1999
	Myosin I	S. cerevisiae (Myo3p, Myo5p) S. pombe (Myo1p)	Cell polarisation	Lechler *et al.*, 2000, Evangelista *et al.*, 2000, Lee *et al.*, 2000,
Class II	Cortactin	Metazoans	Cortical cytoskeleton, endocytosis	Weed *et al.*, 2000, Weaver *et al.*, 2001
	Abp1p	S. carevisiae Metazoans	Endocytosis	Goode *et al.*, 2001
Unknown class	Pan1p/Eps15	S. cerevisiae Metazoans	Endocytosis	Duncan *et al.*, 2001

Table 1.5 Nucleation-promoting factors (NPFs) of the Arp2/3 complex

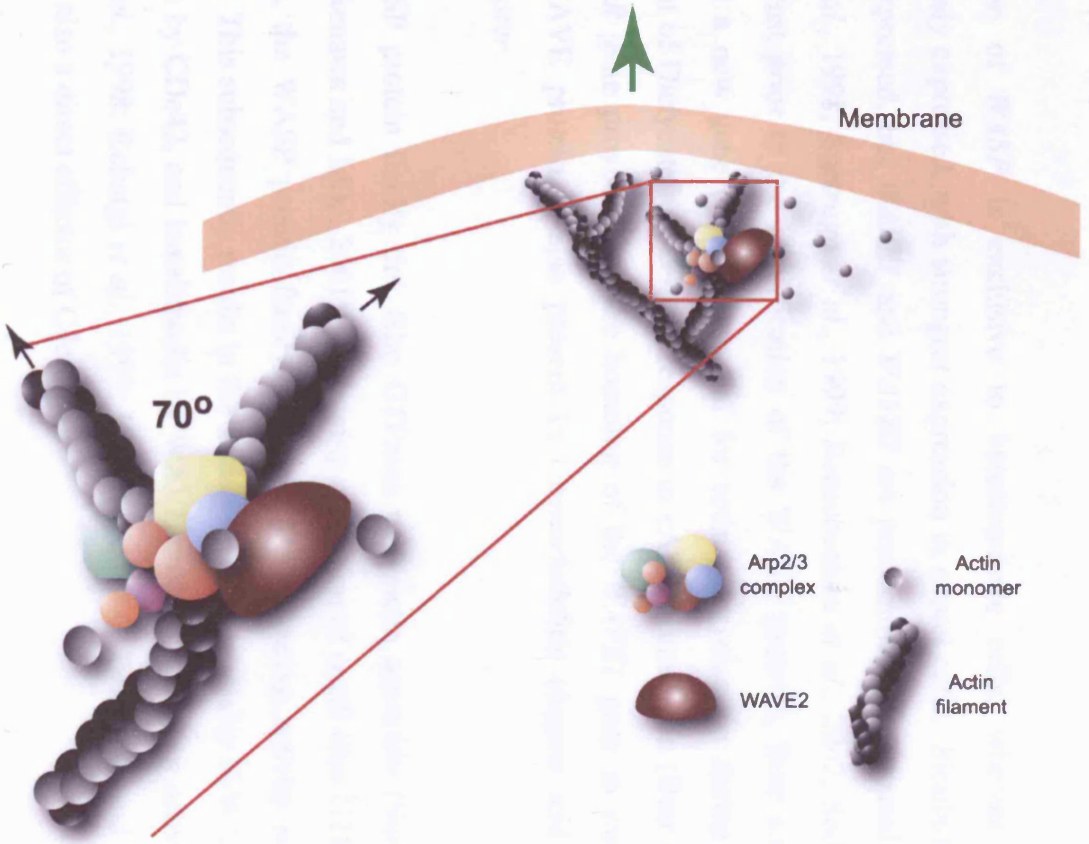


Figure 1.12 Schematic of actin nucleation through activation of the Arp2/3 complex at the leading edge of a motile cell.

Nucleating-promoting factor WAVE2 binds to and activates the Arp2/3 complex leading to formation of Y-shaped branches. Green arrow denotes direction of extending membrane.

1.11 The WASP family of proteins

In mammals the WASP family of proteins is comprised of the founder member WASP, its homologue neural WASP (N-WASP), and three WASP family verprolin homologous (WAVE) proteins (WAVE1-3). WASP was originally identified as the mutated gene in patients with Wiskott-Aldrich syndrome (WAS, Derry *et al.*, 1995). Wiskott-Aldrich syndrome is an X-linked disorder characterised by thrombocytopenia, eczema, and immunodeficiency.

Expression of *WASP* is exclusive to hematopoietic cells, whereas *N-WASP* is ubiquitously expressed, with strongest expression in neural cells. Similarly, *WAVE2* is widely expressed, but *WAVE1* and *WAVE3* are predominantly expressed in the brain (Bear *et al.*, 1998; Suetsugu *et al.*, 1999; Benachenhou *et al.*, 2002; Soderling *et al.*, 2003). Just prior to the identification of the WAVE proteins, Bear and colleagues identified a new gene (*SCAR*) essential for actin reorganisation during chemotactic movement of Dictyostelium cells in response to cAMP stimulation (Bear *et al.*, 1998). The *SCAR* gene proved to be the homolog of the *WAVE1* gene in mammals. A SCAR/WAVE protein is also present in *Caenorhabditis elegans* and *Drosophilia melanogaster*.

The WASP protein family link Rho GTPases with actin assembly (Suetsugu *et al.*, 1999; Takenawa and Miki, 2001). Following activation of small Rho GTPases (Cdc42 and Rac), the WASP protein family stimulate the nucleating activity of the Arp2/3 complex. This subsequently results in the formation of filopodia by N-WASP following activation by CDc42, and lamellipodia by WAVE proteins following activation by Rac (Miki *et al.*, 1998; Rohatgi *et al.*, 1999; Kim *et al.*, 2000; Takenawa and Miki, 2001). WASP is also a direct effector of Cdc42.

The WASP protein family have a common molecular organisation (figure 1.13). All members (WASP, N-WASP, and WAVE1-3) contain a verprolin homology/WASP homolgy (V/WH2), central (C), and acidic (A) domain (VCA/WCA) at the carboxy terminus. In addition, all members have a proline rich region preceded by a basic motif, upstream of the WCA domain. The WASP protein family can then be placed into subcategories based on the sequence at their N-terminus. The WASP-like category

consists of WASP and N-WASP, and the three WAVE proteins compose the remaining category. WASP and N-WASP contain a central GTPase binding domain (GBD) and an N-terminal WASP homology domain 1 (WH1), also referred to as Ena-VASP homology domain 1 (EVH1, Callebaut *et al.*, 1998). These domains (GBD and WH1) are absent from the WAVE proteins. Instead, the WAVEs have a SCAR homology domain (SHD), also referred to as a WAVE homology domain (WHD).

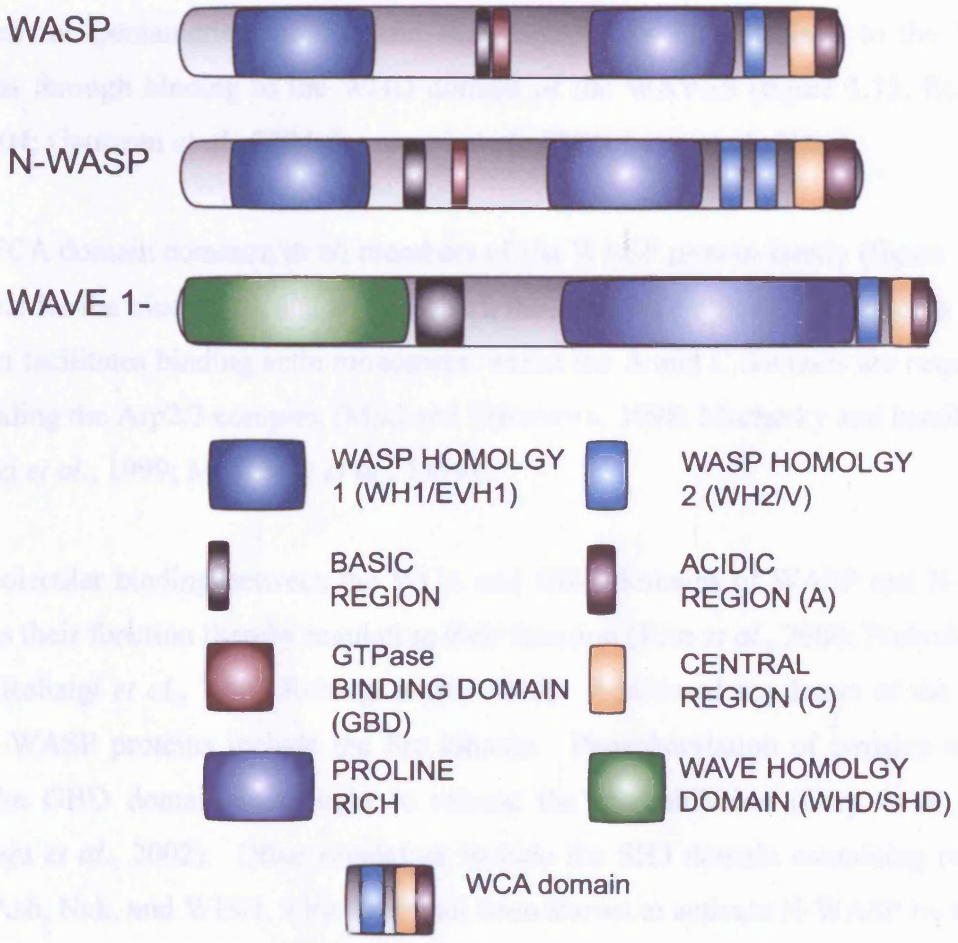


Figure 1.13 Domains within the WASP protein family members.

Note, the WCA domain in N-WASP contains two WH2 domains.

The variation in the domains at the N-terminus of the WASP protein family results in differential regulation of the two subcategories. The GBD domain of WASP and N-WASP binds directly to GTP-Cdc42, whilst the basic region binds PIP2 (figure 1.14, Symons *et al.*, 1996; Miki *et al.*, 1998; Rohatgi *et al.*, 2000; Higgs and Pollard, 2000; Prehoda *et al.*, 2000). In the absence of GTP-Cdc42, WASP and N-WASP are

autoinhibited by an intramolecular interaction between the WCA and GBD domains (Miki *et al.*, 1998; Rohatgi *et al.*, 1999; Higgs and Pollard, 2000; Kim *et al.*, 2000). The WAVE proteins do not directly interact with the small GTPase Rac. Regulation of the activity of WAVES is believed to result from their interaction within a pentameric complex including PIR121 (p53 inducible messenger RNA), Nap1 (Nck-associated protein), Abi (Abl interactor, see section 1.13), and HSPC300 (Eden *et al.*, 2002). PIR121 shares sequence identity with a Rac effector, p140Sra-1 (specifically Rac1-associated protein, Kobayashi *et al.*, 1998), and so is thought to mediate the link between the pentameric complex and Rac. Abi links PIR121/Nap1 to the WAVE proteins through binding to the WHD domain of the WAVES (figure 1.15, Echarri *et al.*, 2004; Gautreau *et al.*, 2004; Innocenti *et al.*, 2004; Leng *et al.*, 2005).

The WCA domain common to all members of the WASP protein family (figure 1.13) is essential for the binding to, and activation of, the Arp2/3 complex (section 1.10). The V domain facilitates binding actin monomers, whilst the A and C domains are responsible for binding the Arp2/3 complex (Miki and Takenawa, 1998; Machesky and Insall, 1998; Rohatgi *et al.*, 1999; Machesky *et al.*, 1999).

Intramolecular binding between the WCA and GBD domains of WASP and N-WASP inhibits their function thereby regulating their function (Kim *et al.*, 2000; Prehoda *et al.*, 2000; Rohatgi *et al.*, 1999; Rohatgi *et al.*, 2000). Additional regulators of the WASP and N-WASP proteins include the Src kinases. Phosphorylation of tyrosine residues near the GBD domain is thought to release the autoinhibition (Cory *et al.*, 2002; Suetsugu *et al.*, 2002). Other regulators include the SH3 domain containing proteins, Grb2/Ash, Nck, and WISH, which have all been shown to activate N-WASP by binding the proline-rich region (Carlier *et al.*, 2000; Fukouka *et al.*, 2001; Rohatgi *et al.*, 2001).

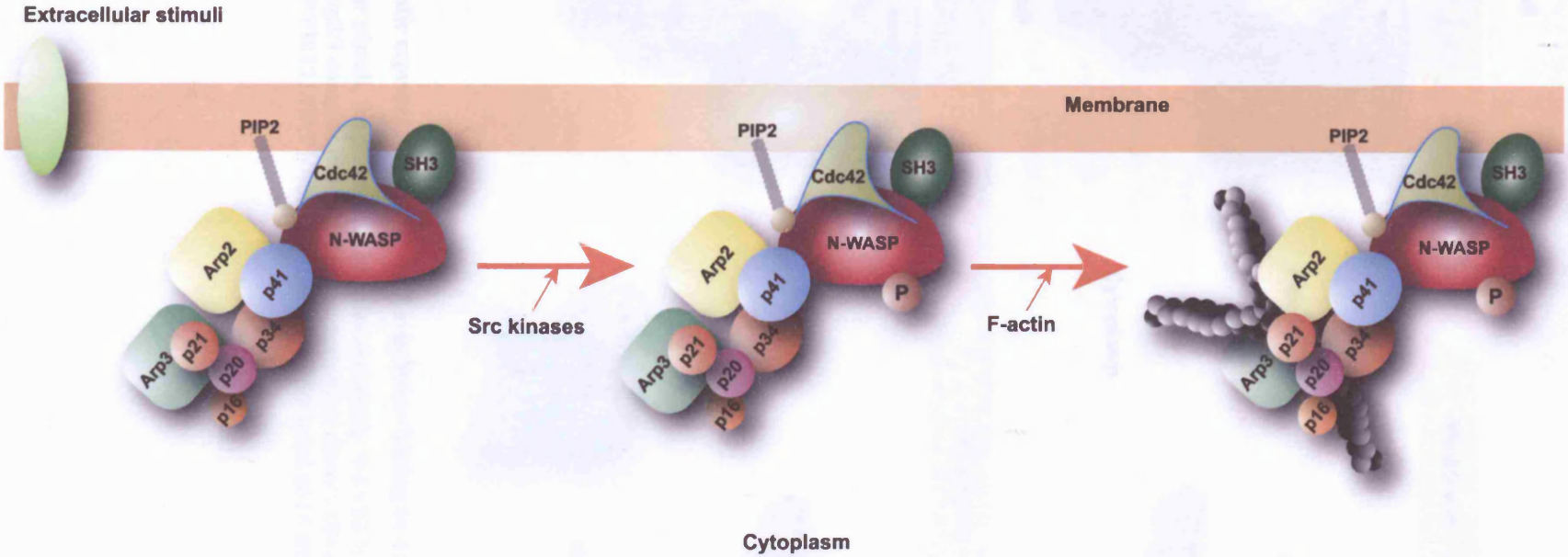


Figure 1.14 Schematic representation of WASP regulation leading to Arp2/3 complex activation

Following extracellular stimuli, the autoinhibition of N-WASP is relieved through the binding of either, Cdc42 and PIP2, an SH3 containing protein. Phosphorylation of tyrosine by Src kinases enhances N-WASPs ability to activate the Arp2/3 complex. Adapted from Bompard and Caron, 2004.

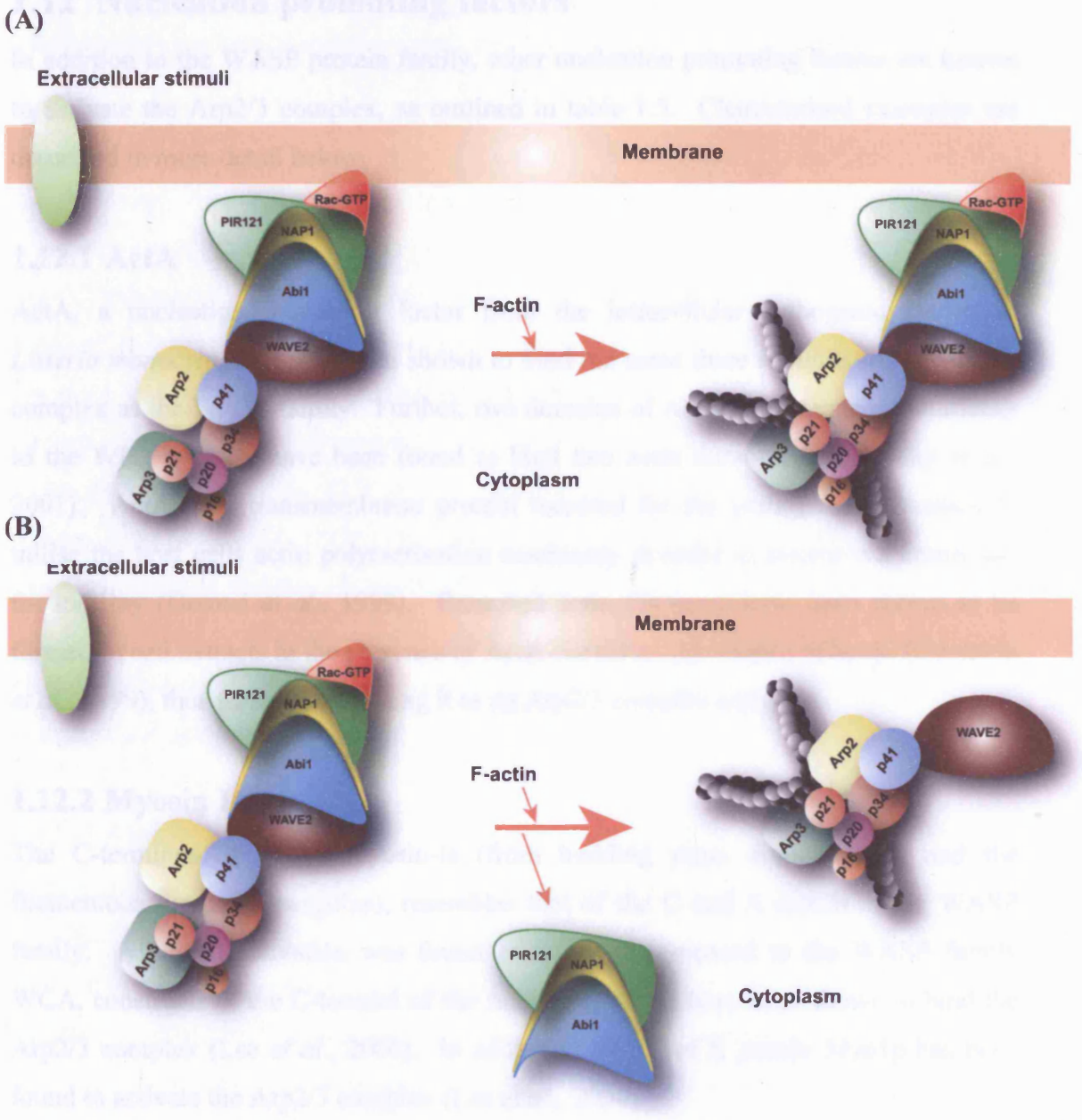


Figure 1.15 Schematic representation of WAVE regulation leading to Arp2/3 complex activation

Following extracellular stimuli, the pentameric complex containing WAVE2 is localised to the membrane and activates the Arp2/3 complex either in the presence (A) or absence (B) of the remainder of Rac-GTP/PIR121/NAP1/Abi1. Adapted from Bompard and Caron, 2004.

1.12 Nucleation promoting factors

In addition to the WASP protein family, other nucleation promoting factors are known to activate the Arp2/3 complex, as outlined in table 1.5. Characterised examples are described in more detail below.

1.12.1 ActA

ActA, a nucleation promoting factor from the intracellular pathogenic bacterium *Listeria monocytogenes*, has been shown to bind the same three subunits of the Arp2/3 complex as the WASP family. Further, two domains of Act A with sequence similarity to the WH2 domain, have been found to bind two actin monomers (Zalevsky *et al.*, 2001). ActA, is a transmembrane protein required for the pathogenic bacterium to utilise the host cells actin polymerisation machinery in order to assemble a comet tail for motility (Dramsi *et al.*, 1998). Branched actin filaments have been shown to be formed in cell extracts in the presence of ActA coated on the surface of beads (Cameron *et al.*, 1999), thus further implicating it as an Arp2/3 complex activator.

1.12.2 Myosin I

The C-terminus of fungal myosin-Is (from budding yeast, fission yeast, and the filamentous fungus Aspergillus), resembles that of the C and A motifs in the WASP family. Although activation was found to be weak compared to the WASP family WCA, constructs of the C-termini of the fungal myosin-Is have been shown to bind the Arp2/3 complex (Lee *et al.*, 2000). In addition, the tail of *S. pombe* Myo1p has been found to activate the Arp2/3 complex (Lee *et al.*, 2000).

1.12.3 Cortactin

Cortactin contains an N-terminal acidic domain and an F-actin binding domain, which have been shown to be necessary to activate the Arp2/3 complex (Weed *et al.*, 2000; Uruno *et al.*, 2001; Weaver *et al.*, 2001). Cortactin binds the Arp2/3 complex via the Arp3 subunit (Weaver *et al.*, 2002) and is localised at the leading edge of lamellipodia in metazoans. Cortactin appears to stabilise the branches formed by the Arp2/3 complex (Weaver *et al.*, 2001), thereby possibly prolonging the lifetime of dendritic networks. Recently, cells in which cortactin had been knocked down, exhibited a defect

in the persistence of lamellipodial protrusions, thus implying that cortactin does have a role in enhancing lamellipodial protrusions (Bryce *et al.*, 2005).

1.12.4 Abp1p

Like cortactin, Abp1p does not contain a G-actin binding domain, but instead has an F-actin binding domain. Abp1p was the first actin associated protein to be identified in yeast (Drubin *et al.*, 1990) and consist of an ADF/cofilin homology domain (ADF/F-H), a helical region, a proline-rich domain, and an SH3 domain. Overexpression of Abp1p leads to severe defects in cellular actin organisation (Drubin *et al.*, 1990), which corresponds with the localisation of Abp1p to cortical actin patches (Drubin *et al.*, 1990). Activation of the GTPase Rac results in the translocation of Abp1p to the leading edge of cells (Kessels *et al.*, 2000). Abp1p contains two acidic motifs, which are essential for activation of the Arp2/3 complex (Goode *et al.*, 2001). The ADF-H domain mediates F-actin binding (Goode *et al.*, 2001). Binding to F-actin filaments and to the Arp2/3 complex, enables Abp1p to recruit the Arp2/3 complex to the sides of filaments and possibly link endocytosis with actin filament dynamics.

1.12.5 RickA

RickA was recently identified as a protein expressed on the surface of *Rickettsia conorii*, which activates the Arp2/3 complex leading to the generation of long-unbranched actin filaments (Gouin *et al.*, 2004). The protein sequence of RickA contains a central proline-rich region, an N-terminal G-actin binding domain, and a C-terminal CA domain. In vitro studies of the RickA protein confirmed activation of the Arp2/3 complex, but with less efficiency than ActA (see section 1.12.1), and expression of RickA at the inner face of the plasma membrane in mammalian cells, was found to result in the formation of filopodia (Gouin *et al.*, 2004). RickA is the most closely related bacterial nucleation promoting factor to the WASP protein family.

1.13 The Abelson (Abl)-interactor (Abi) family of proteins

First identified as a substrate and a binding partner for Abelson tyrosine kinases (Dai and Pendergast, 1995; Shi *et al.*, 1995; Juang and Hoffmann, 1999), the Abi proteins are instrumental in the intracellular reorganisation of actin through the regulation of Rac-dependent pathways. Members of the Abi family include mammalian Abi1, Abi2, and Abi3/NESH (New molecule including SH3), *Drosophila melanogaster* Abi (dAbi), *Xenpous laevis* Xlan4 (Xenopus laevis animal 4), *Dictyostelium discoideum* Abi, and *Caenorhabditis elegans* Abi (Reddy *et al.*, 1992; Dai and Pendergast; 1995; Shi *et al.*, 1995; Juang and Hoffmann, 1999; Miyazaki *et al.*, 2000).

The amino- and carboxy-terminal regions of the Abi proteins are well conserved across species, whilst the middle portion of the molecule varies considerably (Echarri *et al.*, 2004). Conserved domains amongst the Abi proteins include an SH3 domain within the carboxy terminus, a homeodomain homologous region, and SNARE and WAVE-binding (WAB) domains at the amino terminus (Dai and Pendergast; 1995; Shi *et al.*, 1995; Stradal *et al.*, 2001; Echarri *et al.*, 2004).

Abi proteins are localised to sites of actin polymerisation in protrusive membrane structures in particular; Abi1 and Abi2 have been shown to localise to the leading edge of lamellipodia and filopodia in highly motile B16 melanoma cells (Stradal *et al.*, 2001). The cellular localisation pattern observed for Abi1 and Abi2 is comparable to the cellular localisation pattern observed with the WAVE proteins (Hahne *et al.*, 2001, Stradal *et al.*, 2001). Two recent reports showed Abi1 to regulate the localisation of WAVE1 and WAVE2 to the leading edge of lamellipodia (Echarri *et al.*, 2004; Leng *et al.*, 2005). In addition WAVE1 levels were also found to be regulated by Abi1 (Echarri *et al.*, 2004).

Abi1 and Abi2 act as scaffolding proteins, leading the assembly of various multi-molecular complexes (Innocenti *et al.*, 2003; Stradal *et al.*, 2004; Disanza *et al.*, 2004). One such complex is the WAVE2-Abi1-NAP1-PIR121 complex (figure 1.16; Innocenti *et al.*, 2004). Abi1 directly binds WAVE2 forming the platform on which NAP1 and PIR121 associate (figure 1.16; Innocenti *et al.*, 2004). Abi2 has been reported to form a complex with WAVE1, Nap-1, HSPC300, and PIR121 (Eden *et al.*,

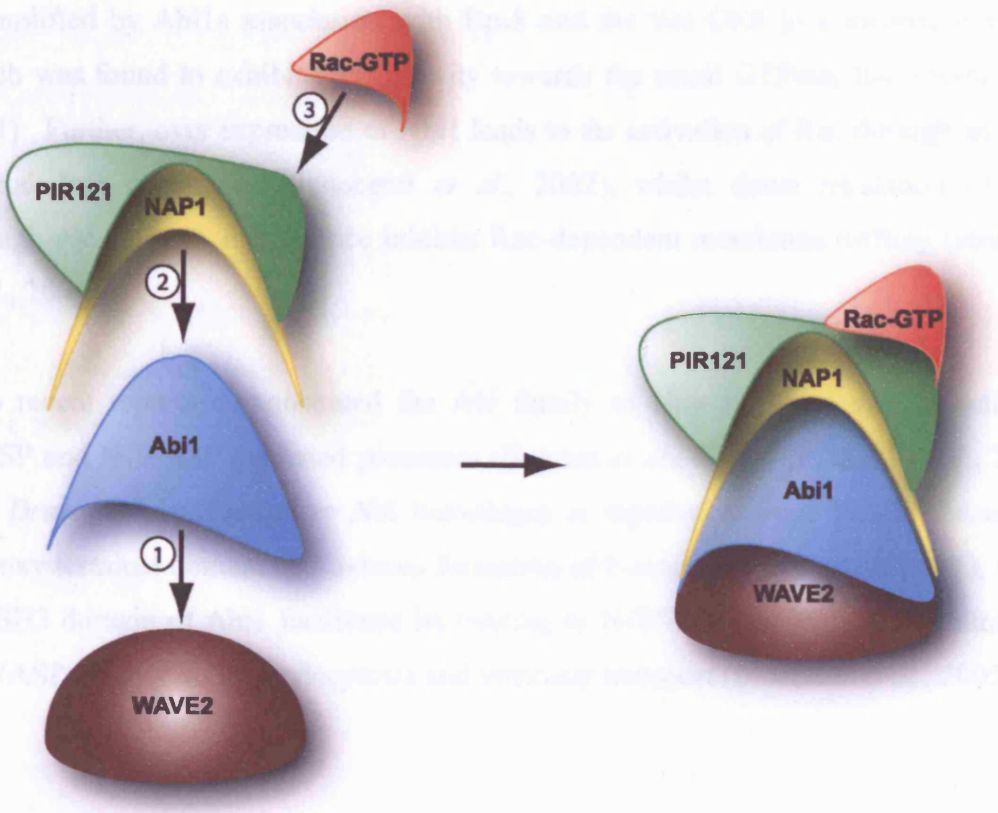


Figure 1.16 Abi1 directly binds WAVE2 forming the platform on which NAP1 and PIR121 associate

2002). Other known interactors of the Abi proteins include the adapter protein Eps8, Spectrin, the Sos guanine nucleotide exchange factor, and Grb4 (Biesova *et al.*, 1997; Ziemnicka-Kotula *et al.*, 1998; Scita *et al.*, 1999; Fan and Goff, 2000; Cowan and Henkemeyer, 2001).

The Abi family of proteins link to the signalling pathway, involving the Rac GTPase, is exemplified by Abi1s association with Eps8 and the Sos GEF in a trimeric complex, which was found to exhibit GEF activity towards the small GTPase, Rac (Scita *et al.*, 2001). Further, over expression of Abi1 leads to the activation of Rac through an Eps8-dependedent mechanism (Innocenti *et al.*, 2002), whilst down regulation of Abi1 through use of RNA interference inhibits Rac-dependent membrane ruffling (Innocenti *et al.*, 2004).

Two recent reports demonstrated the Abi family to play a role in the regulation of WASP and N-WASP mediated processes (Bogdan *et al.*, 2005; Innocenti *et al.*, 2005). The *Drosophila melanogaster* Abi homologue is reported to bind WASP through its carboxy terminal domain and induces formation of F-actin (Bogdan *et al.*, 2005), whilst the SH3 domain of Abi1 facilitates its binding to N-WASP resulting in its attuning of N-WASP activity during endocytosis and vesicular transport (Innocenti *et al.*, 2005).

1.14 Aims of this thesis

At the outset of this study, a family with X-linked congenital cataract (CXN) had been mapped to the short arm of the human X chromosome (Francis *et al.*, 2002). The CXN locus resided within the locus for Nance-Horan syndrome (NHS) for which a similar cataract phenotype in affected males is seen. It had therefore been previously hypothesised that X-linked congenital cataract and Nance-Horan syndrome were possibly allelic. In strength of this hypothesis, an X-linked congenital cataract locus in the mouse had been mapped to the distal end of the mouse X chromosome (Stambolian *et al.*, 1994), syntenic with the portion of the human X chromosome for which CXN and NHS had been mapped. Thus, the aim of this thesis was to identify the disease causative gene(s) in the CXN family and families with NHS, through positional cloning. Following identification of the causative gene(s) for CXN and NHS, functional analyses of the protein product(s) would be initiated to unravel the mechanisms of lens development and cataract formation.

- Chapter Two -

Materials and Methods

2.1 Materials

All reagents were purchased from ABgene, UK, Amersham Biosciences, UK, Invitrogen, UK, Promega, UK, and Sigma, UK, and QIAGEN, UK, unless otherwise stated.

2.1.1 Reagents and Buffers

2.1.1.1 Tissue culture reagents

All tissue culture reagents were purchased from Invitrogen (Paisley, UK; see table 2.5).

Phosphate buffered saline (PBS): 10 mM phosphate, 0.9 % (w/v) NaCl, pH 7.4

2.1.1.2 Plasmid DNA

X-gal: 20 mg/ml stock

IPTG: 100 mM stock

LB broth: 1 % (w/v) Tryptone, 1 % (w/v) NaCl, 0.5 % (w/v) Yeast extract. Made up with distilled water and autoclaved before use.

LB agar: LB broth with 1.5 % (w/v) agar, autoclaved prior to use.

2.1.1.3 Protein purification, electrophoresis, and Western blotting reagents

Coomassie stain: 0.2 % (w/v) Coomassie blue, 50 % (v/v) methanol, 10 % (v/v) acetic acid.

Coomassie de-stain: 45 % (v/v) methanol, 10 % (v/v) acetic acid.

Cells and tissue homogenisation buffer: 100 mM Tris-HCl, 2 % (v/v) glycerol, 1 % (w/v) SDS, 1 % (w/v) 2-β mercaptoethanol

Ponceau S stain: 3 % (v/v) Trichloroacetic acid, 0.2 % (w/v) Ponceau S

SDS-PAGE sample buffer (5×): 125 mM Tris-HCl (pH 6.8), 4 % (v/v) SDS, 20 % (v/v) glycerol, 10 % (v/v) 2-mercaptoethanol and 0.2 % (w/v) bromophenol blue.

SDS-PAGE running buffer (1×): 25 mM Tris-HCl, 190 mM glycine, 3.5 mM SDS.

Tris-acetate EDTA buffer (TAE, 1×): 40 mM Tris-HCl, 40 mM acetic acid, 1 mM EDTA

Western blotting transfer buffer (1×): 25 mM Tris-HCl, 190 mM glycine, 20 % (v/v) methanol

Western blotting blocking buffer: 5 % (w/v) "Marvel" milk powder, 0.05 % (v/v) Tween-20, in 1 × PBS.

Western blotting wash buffer: 0.05 % (v/v) Tween-20 in 1 × PBS.

Lysate buffer: 150 mM NaCl, 10 mM sodium phosphate (pH 7.0), 1 % (v/v) Triton X-1:100 dilution of protease inhibitor cocktail, in 1 × PBS.

2.1.1.4 Actin assay reagents

General actin buffer: 5 mM Tris-HCl (pH 8.0), 0.2 mM $CaCl_2$, 0.2 mM ATP, 0.5 mM DTT

Actin polymerisation buffer (10 ×): 100 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$, 10 mM ATP, and 0.5 M KCl.

2.1.1.5 General reagents

1 M Tris-HCl (pH 7.5, or 8): adjusted pH to 7.5 or 8.0 with 5 M NaOH, autoclaved before use.

5 M NaOH: prepared with SDW.

10 % (w/v) SDS: prepared with SDW.

Unless otherwise stated, room temperature was approximately 21-23°C.

2.2 Molecular Biology Techniques

2.2.1 Polymerase Chain Reaction (PCR)

A typical PCR mix (25 µl) consisted of 50-150 ng DNA, 2 pmol of each primer, 200 µM of each dNTP, 1.5 - 3 mM $MgCl_2$, 1 × NH_4 reaction buffer, 0.5 U BIOTAQ-polymerase (Bioline). PCRs were carried out on Techne Genius PCR thermocylers equipped with heated lids thus eliminating the necessity to overlay PCR reactions with mineral oil to prevent evaporation. Cycling conditions consisted of an initial denaturation step of 95°C for 5 minutes followed by 30 to 40 cycles of 95°C for 30 seconds, 52°C – 64°C (depending on the required annealing temperature) for 30 seconds and 72°C for 30 seconds (extension time varied depending on the size of the amplicon). A final cycle of 72°C for 5 minutes completed the extension.

In addition to the standard PCR mix mentioned above, a 2 × ReddyMix™ PCR Master Mix (Abgene, Epsom, UK) was used. The Abgene master mix consisted of 1.25 U Thermoprime Plus DNA Polymerase, 75 mM Tris-HCl (pH 8.8 at 25°C), 20 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.01% (v/v) Tween® 20, 0.2 mM each of dATP, dCTP, dGTP and dTTP, precipitant and red dye (for electrophoresis), 2 pmol of each primer and 5-150 ng DNA. For templates that were difficult to amplify, the Abgene Extensor Hi-Fidelity PCR Master Mix was used and consisted of consisting of 1.25 U Thermoprime Plus DNA Polymerase, 2.25 mM $MgCl_2$, 500 µM each of dATP, dCTP, dGTP, and dTTP.

2.2.2 Primer design and determination of annealing temperatures

Annealing temperatures for newly designed primer pairs were approximated from their melting temperatures (T_m). To calculate the T_m, which is dependent on a primer nucleotide sequence, the following formula was used:

$$4(G + C) + 2(A+T) = T_m$$

Equation 1 Formula for calculating the melting temperature of primers

By lowering the value obtained by 2-5°C, an annealing temperature could be assigned (e.g. a T_m of 60°C would indicate an annealing temperature of 57-58°C).

In order to design primer pairs for PCR amplification, several rules were followed: random base distribution and similar GC content for both primers, an anchoring C or G at the 3' end of each primer, minimal secondary structure (i.e. self-complementary) and low complementarity to each other, especially in the 3' region as to avoid the incidence of 'primer dimer' formation, no greater than 4°C difference between the T_m values of both primers and, bearing in mind the above constraints, a primer length of at least 20 nucleotides- to increase the sequence specificity.

2.2.3 Reverse Transcription (RT) PCR

Synthesis of cDNA from RNA extracted from tissue was performed using M-MLV Reverse Transcriptase (Promega, Southampton, UK). First strand cDNA synthesis reactions (20 µl) consisted of M-MLV Reverse Transcriptase 5 × buffer, 200 µM of each dNTP, 0.5 ug Oligo $(dT)_{15}$ primer, 20 U RNasin® Rnase inhibitor, 100 U M-MLV Reverse Transcriptase Rnase H Minus, and 2-3 µg of total RNA.

Reverse transcription was performed in a Techne Genius PCR thermocyler at 55 °C for 1 hour followed by 10 minutes at 65 °C. The double stranded cDNA was then either used immediately as a template for PCR or stored at –20°C.

2.2.4 Colony PCR

Identification of colony forming units transformed with the required vector and insert was carried out using colony PCR. Following plating out of transformed bacterial cultures and incubation overnight at 37 °C, white forming colonies were picked using a sterile pipette tip and used as a template for PCR (section 2.2.1). The sterile tip containing the picked colony was dipped in a 0.2 ml eppendorf tube containing 25 µl of PCR mix and then stored for subsequent propagation. PCRs were carried out on Techne Genius PCR thermocylers.

2.2.5 Size Fractionation of DNA

2.2.5.1 Agarose gel electrophoresis

PCR products, vectors, clones, and enzyme digests were size-fractionated on an appropriate % agarose gel depending on the fragment size (0.8 - 1 % for 2 - 20 Kb fragments, 1 - 2 % for 1 - 2 Kb fragments and 3 % for 70 bp – 1 Kb fragments). Agarose gels were prepared by weighing out the appropriate amount of agarose (% w/v) in a conical flask and adding the correct volume of 1 × TAE. The flask contents were mixed by gentle swirling and heated in a microwave to dissolve the agarose. The dissolved agarose was left to cool to approximately 60 °C before the addition of 5 µl of 0.5µg/ml ethidium bromide. The melted agarose was then poured into an appropriate pre-sealed casting plate containing a well comb. Agarose gels were left to set for 20-30 minutes at room temperature.

Generally, 5 µl of PCR product or DNA was mixed with 1 µl blue/orange 6 × gel loading dye (Promega, Southamton, UK) and loaded into the well (except for where 2 × ReddyMix™ PCR Master Mix was used). An appropriate DNA size marker was also loaded, depending on the anticipated size of the DNA fragments (see table 2.1).

ϕX174/HaeIII	*1Kb DNA ladder*
1,358	10,000
1,078	8,000
872	6,000
602	5,000
310	4,000
281	3,000
271	2,500
234	2,000
194	1,500
118	1,000
-	750
-	500
-	250

Table 2.1 DNA marker sizes (bp) for, ϕX174/HaeIII and 1Kb DNA ladder

2.2.6 QIAquick Gel Extraction

Inserts digested out from vectors, or PCR products were electrophoresed on a 0.5 % (w/v) agarose gel, gel extracted, and purified. DNA in the range of 70 bp to 10 Kb was extracted and purified using the QIAquick gel extraction kit according to the manufacturers guidelines.

Upon resolving DNA on an agarose gel, the band of interest was excised from the gel and the gel slice weighed. Three volumes of buffer QG was then added to one volume of gel (100 mg is equivalent to 100 µl) and incubated at 50 °C for ten minutes with gentle vortexing every two to three minutes. For DNA fragments outside the range of 500 bp to 4 Kb, one gel volume of isopropanol was added to the sample and mixed. A QIAquick spin column was then placed in a 2 ml collection tube and the DNA bound to the membrane by applying the sample to the column and centrifuging for one minute at approximately 17,900 × g. The flow through was discarded and 750 µl of buffer PE added and centrifuged for one minute at approximately 17,900 × g. The flow though was then discarded and an additional centrifuge step performed for one minute at approximately 17,900 × g in order to remove any residual buffer. The column was then transferred to a sterile 1.5 ml microcentrifuge tube. 50 µl of sterile distilled water was then applied to the centre of the QIAquick membrane and centrifuged for one minute at approximately 17,900 × g to elute the DNA.

2.2.7 Genotyping

Microsatellite markers were selected and primer sequences obtained from the genome database (http://www.gdb.org). Novel microsatellite markers were identified using an in-house nucleotide repeat locator (World-wide Webster web designs 2000) and available genomic sequence. Primers for microsatellite markers were labelled with either 6-FAM (blue) or HEX (green) probes (Genosys Biotechnologies Ltd., Cambridgeshire, UK). Microsatellites were amplified from patient genomic DNA by standard PCR (section 2.2.1) with appropriate annealing temperatures (table 3.2). PCR products were then loaded and analysed on an ABI 3100 genotyper according to manufacturers instructions (Applied Biosystems). Genotypes were determined in a double-blind manner.

2.2.8 WAVE® Nucleic Acid Fragment Analysis System

Denaturing high-performance liquid chromatography (DHPLC) allows the automated detection of single base substitutions, insertions and deletions. Their detection is based around the formation and separation of double-stranded DNA fragments containing mismatched bases (heteroduplexes). The WAVE system (Transgenomic, Inc., San Jose, CA) allows the separation of heteroduplexes from homoduplexes in the absence of radioactivity, acrylamide gels or chemical denaturants. Mismatched bases within a heteroduplex begin to separate at a critical temperature whilst the matched bases within homoduplexes remain intact. Consequently, a lower percentage of organic mobile phase (acetonitrile) is required to disrupt the interactions between heteroduplexes and the column matrix as opposed to the interactions between homoduplexes and the column matrix. Heteroduplex DNA subsequently elutes earlier in the gradient and is detected by additional peaks in the chromatogram. Additional peaks within the chromatogram compared to a reference "wild type" DNA amplicon indicate the presence of a sequence variant.

A 6 µl volume of PCR product was made up to 10 µl with H_2O and subjected to an additional 5 min 94°C denaturing step followed by gradual reannealing from 94°C to 65°C over a 30 minute period to allow formation of heteroduplexes. PCR amplicons were then loaded (5 µl) on a DNASep column packed with C_{18} alkylated, polystyrene-divinylbenzene polymeric beads (Transgenomic Inc., San Jose, CA). The appropriate temperature for resolving heteroduplex molecules was calculated using the DHPLC melting algorithm. Hetero- and homo-duplexes were eluted with a linear acetonitrile gradient at a flow rate of 0.9 ml/min and detected at 260 nm.

2.2.9 RNA extraction

RNA was extracted from various mouse tissues using the Ambion Totally RNA kit. Samples were weighed and homogenised in 10 volumes of denaturation solution using an electronic homogeniser. An equal volume of Phenol:Chloroform:IAA was added to the lysate (starting volume), vortexed vigorously for 1 minute and stored on ice for 10 minutes. The suspension was then centrifuged at 4°C for 15 minutes at 17,900 × g. The upper aqueous phase was transferred to a new vessel and 1/10 the aqueous phase

volume of sodium acetate solution was added and inverted for about 10 seconds. One starting volume of Acid-Phenol:Chloroform was then added and vortexed vigorously for 1 minute. The resulting suspension was stored on ice for 10 minutes and centrifuged at 4 °C for 15 minutes at 17,900 × g. The upper aqueous phase was transferred to a new RNase-free vessel and an equal volume of isopropanol added, and mixed thoroughly. The preparation was then stored at –20 °C for 30 minutes. The precipitation mixture was centrifuged at 4 °C for 15 minutes at 17,900 × g and the supernatant carefully removed and discarded. The RNA pellet was then washed in 200 µl of 70 % ethanol, gently vortexed, followed by centrifugation at 4 °C for 5 minutes at 9,000 × g. The ethanol supernatant was removed and discarded and the RNA pellet resuspended in 30 - 50 µl of DEPC water/EDTA. Resuspended pellets were stored at –80 °C until further use.

2.2.10 Automated DNA sequencing

Both, PCR products and cloned DNA fragments were sequenced using ABI BigDye terminator cycle sequencing chemistry version 3.1 on an ABI 3100 automated sequencer (Applied Biosystems). Before sequencing, PCR products were purified, removing unincorporated primers and dNTPs. An aliquot of the amplification product (2 µl) was purified by the addition of 23 µl dH_2O, 1 U shrimp alkaline phosphatase (SAP, Amersham LifeScience, Buckinghamshire, UK) and 1 U Exonuclease I (United States Biochemical, Ohio, USA) in SAP buffer, and incubated at 37 °C for 15 minutes followed by 80 °C for 15 minutes. A 13 µl aliquot of purified PCR product was then used for cycle sequencing. Cycle sequencing reactions (20 µl) consisted of 1 µl BigDye, 1 µl of 1 pmol sequence specific primer, 13 µl of purified PCR product, and 5 µl TM buffer (200 mM Tris.HCl pH 9.0, 5 mM $MgCl_2$). For cycle sequencing of cloned DNA fragments, 2 µl of (100-250 ng) cloned DNA, 2 µl of 1 pmol sequence specific primer, 2 µl BigDye and 3 µl TM buffer was made up to a total volume of 20 µl. Cycle sequencing reactions (30 cycles of, 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes) were carried out in a Techne Genius PCR thermocyler. The DNA was then precipitated by addition of 26 µl precipitation solution (50 ml 95% (v/v) Ethanol, 2 ml 3 M NaOAc pH 5.2), vortexed and left on ice for 10 minutes before centrifugation at 17,900 × g for 30 minutes. The supernatant was then aspirated and 200

µl of 70 % ethanol was added to the DNA pellet and centrifuged at 17,900 × g for 10 minutes. The supernatant was removed and the tubes heated (with the lids open) at 65°C for 1-5 minutes. The DNA pellet was then resuspended in 11 µl formamide before being loaded onto the ABI 3100 automated sequencer.

2.2.11 Restriction enzyme digest

Single and double digests of mini- and maxi-prepped plasmid DNA was carried out in a total volume of 10 µl. Restriction enzyme digests consisted of 2 µl of mini/maxi-prepped DNA, 1 µl of 12 U/µl restriction enzyme for single digests or 0.5 µl of each enzyme for double digests, and 1 µl 10 × Buffer (Promega, Southampton, UK). Digests were incubated for 2 hours at 37 °C in a Techne Genius PCR thermocyler and then resolved on a 0.5 % (w/v) agarose gel (section 2.2.5.1).

2.2.12 Quantification of DNA and RNA

Following DNA/RNA purification, the concentration was determined by measuring absorbance at 260 nm on a spectrophotometer. 5 µl of DNA was made up to 1 ml with sterile water. The sample was then placed in a cuvette and absorbance read at 260 nm. An absorption of 1 OD (A) is equivalent to 50 µg/ml dsDNA, 33 µg/ml ssDNA, or 40 µg/ml RNA. These standards were used to calculate the concentration of DNA, factoring in the dilution factor. To check for purity, DNA absorption was also measured at 280 nm. The ratio $A_{[260]}/A_{[280]}$ is used as an estimate of the purity of nucleic acids as contaminating proteins absorb at 280 nm. Pure DNA has a ratio of approximately 1.8 and pure RNA, a ratio of approximately 2.0.

2.3 DNA Plasmid Preparation

2.3.1 Ligation reaction

Digested DNA fragments were purified (section 2.2.6) and ligated into the appropriate plasmid vector pre-digested with the same sites, in a 3:1 molar ratio. The ligation reaction contained 2 × rapid ligation buffer (Promega, Southampton, UK), approximately 500 ng of plasmid vector DNA, 3 Weiss units of T4 DNA ligase

(Promega, Southampton, UK) and x ng of insert DNA, in a total volume of 20 µl. Ligation reactions were incubated at 4 °C overnight.

2.3.2 Transformation of competent cells

JM109 high efficiency competent cells (Promega) were thawed on ice and gently mixed. 50 µl of JM109 cells were then transferred to a 1.5 ml eppendorf tube containing 5 µl aliquot of the ligation reaction, and gently mixed. The cells were then heat shocked for 45-50 seconds at 42°C and immediately returned to ice. 950 µl of room temperature SOC medium (Invitrogen) was added and the cells incubated for 1.5 hours at 37°C with shaking at 200 rpm. 100 µl of the transformation culture was plated onto an LB/ampicillin plate and incubated overnight (16-24 hours) at 37 °C.

Single colonies were picked and used to inoculate 5 ml LB broth (containing the appropriate antibiotic) and grown overnight with vigorous shaking at 37 °C. Plasmid DNA was then harvested from the overnight cultures as described in section 2.3.3 and digested with appropriate restriction enzymes to confirm the presence of an insert. Minipreps containing inserts were sequence verified to check for PCR induced errors before being stored at - 20 °C. 750 µl of the overnight culture was mixed with an equal volume of 50 % (v/v) glycerol and stored at –80°C as glycerol stocks.

2.3.3 Purification of plasmid DNA

Plasmid DNA was extracted and purified from bacterial cells using either the Genelute plasmid mini-prep kit (Sigma) or the HiSpeed Plasmid Maxi kit (QIAGEN).

2.3.3.1 Purification of plasmid DNA using the Genelute Plasmid Mini-Prep kit

Cells were pelleted by centrifuging 1.5 ml of overnight culture at 12,000 × g and the supernatant discarded. Cells were resuspended in 200 µl resuspension solution and vortexed. 200 µl of lysis solution was added to the resuspended cells and the tube gently inverted 4 – 6 times to mix. The lysed cell suspension was left to clear for approximately 4 minutes. 350 µl of neutralization solution was then added, the tube inverted 4-6 times and centrifuged at 17,900 × g for 10 minutes. To a binding column,

500 ml of column preparation solution was applied, centrifuged at 17,900 × g for 1 minute, and the flow-through discarded. The clear lysate was then transferred into the binding column and centrifuged for 1 minute at 17,900 × g. The flow-through was discarded and 750 µl of wash solution was added to the binding column and centrifuged for 1 minute at 17,900 × g. The flow-through was discarded and the column centrifuged for an additional 1 minute at 17,900 × g to remove excess flow through. The column was then transferred to a clean sterile 1.5 ml microcentrifuge tube and the DNA eluted by the addition of 50 µl of water and a final centrifugation step at 17,900 × g for 1 minute.

2.3.3.2 Purification of plasmid DNA using the HiSpeed Plasmid Maxi kit

A single colony from a freshly streaked selective plate was picked and used to inoculate a starter culture of 5 ml LB medium (see section 2.3.2) containing the appropriate antibiotic. The inoculated culture was incubated overnight at 37 °C with vigorous shaking (200 rpm). The starter culture was then diluted 1/1000 into 250 ml medium containing the appropriate antibiotic, and grown overnight at 37 °C with vigorous shaking (200 rpm). Overnight cultures were then centrifuged at 6000 × g for 15 minutes at 4 °C to pellet the cells. The bacterial pellet was then resuspended in 10 ml buffer P1, to which 10 ml of lysis buffer P2 was added, mixed gently and incubated at approximately 22 °C for 5 minutes. 10 ml of chilled neutralisation buffer P3 was then added to the lysate, mixed immediately and transferred to a QIAfilter Cartridge. The lysate was then filtered through the QIAfilter Cartridge into a HiSpeed Tip previously equilibrated with 10 ml QBT buffer. The cleared lysate was allowed to enter the resin by gravity flow and was washed with 60 ml buffer QC. The DNA was then eluted with 15 ml buffer QF. The eluted DNA was precipitated by incubating with 10.5 ml isopropanol at approximately 22 °C for 5 minutes. The eluate/isopropanol mixture was transferred to a 30 ml syringe connected to a QIAprecipitator. The eluate/isopropanol mixture was then filtered throught the QIAprecipitator. The bound DNA was washed by passing 2 ml 70 % (v/v) ethanol through the QIAprecipitator. The QIAprecipitator membrane was dried twice, by passing air through it quickly, and attached to a new 5 ml syringe. 1 ml of sterile water was passed through the QIAprecipitator into a sterile 1.5 ml microcentrifuge tube. The eluate was then passed through the QIAprecipitator a second time to ensure maximum yield.

2.3.4 Cloning

2.3.4.1 TA cloning into pGEM-T Easy

Constructs to be TA cloned were PCR amplified (section 2.2.1) using template specific primers, resolved by gel electrophoresis (section 2.2.5) and purified (section 2.2.6). Purified inserts were then ligated (section 2.3.1) into the pGEM-T Easy vector (Promega,), transformed into JM109 electro competent cells (section 2.3.2) and grown overnight on IPTG/X-gal/AMP LB plates. White colonies were selected and the plasmid DNA harvested (section 2.3.3.1). Clones were sequenced (section 2.2.10) to check for PCR introduced errors using vector and insert primers.

2.3.4.2 Generating GST-tagged proteins

The C-terminus of human NHS was directly cloned into pGEX-2T for characterisation of the C-terminal peptide antibody. A 607 bp fragment was amplified using primers engineered with restriction enzyme sites (see table 2.2). Both the PCR product and pGEX-2T vector were digested (section 2.2.11) with the appropriate restriction enzymes (see tables 2.2 and 2.3), ligated and transformed into JM109 electro competent cells (section 2.31 and 2.3.2). Clones were sequenced to check for PCR introduced errors using vector and insert primers.

Primer name	*Sequence 5'-3'*	*Engineered restriction enzyme site*
NHSC1 F	CGGGATCCTCCAAGAGGAAAGTACTTGG	BamHI
NHSC1 R	CGGAATTCTCTATGTTGAACTCTGGGAG	EcoRI
NHSC2 F	CGGGATCCACCATGTTTACTCCTGCAGTGAGCAG	BamHI
NHSC2 R	CGGAATTCTGGTACAGGCAGTCTATGTTGAACTC	EcoRI
NHS-CA1 F	CGGGATCCAGCAGCTCCACGTGCCC	BamHI
NHS-CA1 R	CGGAATTCGTAATTCCACTCACTGTCAC	EcoRI
NHS-CA2 F	CGGGATCCGTTGGCGCTAAACCCTCAG	EcoRI
NHS-CA2 R	CGGAATTCCCCACTCCAGTTGCCACT	BamHI

Table 2.2 Primer sequences used to clone various NHS fragments into prokaryotic and mammalian expression vectors.

For individual expression vectors into which the above fragments were cloned into refer to table 2.3.

2.3.4.3 Generating Myc-tagged proteins

The NHSC2 construct of the human NHS protein (see figure 6.7), was amplified using primers engineered with restriction enzyme sites (see table 2.2). NHSC2 spanned exon 6 to exon 8 of the *NHS* gene and threrefore, was amplified from a mRNA (accession number CR936788) obtained from RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH (Germany).

Amplicons were TA cloned into the pGEM-T Easy vector (section 2.3.4.1) and sequenced using vector and internal primers to check for PCR introduced errors. Following confirmation of the correct sequence, inserts were digested out of the pGEM-T Easy vector (section 2.2.11), sized fractionated on a 0.5 % (w/v) agarose gel (section 2.2.5.1) from which the relevant insert was excised and gel purified (see section 2.2.6). Purified inserts were then ligated into pCMV-Tag3b (Stratagene, UK section 2.3.1) previously digested with the appropriate restriction enzymes (see table 2.3).

Vector	*Ligated with*	*Digested with restriction enzymes -*	*Buffer*	*Incubation temperature (°C)*	*Incubation period (Hours)*
pCMV-Tag3b	NHSC2	BamHI and EcoRI	E	37	2
pGEX-2T	NHSC1	BamHI and EcoRI	E	37	2
pGEX-2T	NHSC2	BamHI and EcoRI	E	37	2

Table 2.3 Vectors, restriction enzymes, and conditions used for subcloning of relevant PCR generated fragments.

Note, all restriction enzymes and buffers used were obtained from Promega (UK) except where indicated (*) in which case the restriction enzymes and buffers were obtained from Amersham Biosciences (UK).

2.4 Protein Based Techniques

2.4.1 Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

2.4.1.1 Resolving protein samples by SDS-PAGE

Protein samples were analysed by SDS-PAGE using the Mini-PROTEAN II electrophoresis cell (Bio-Rad, Hemel Hempstead, Hertfordshire). Casting gels had dimensions of 9 cm × 7 cm, and were 0.75 mm thick. Samples were loaded onto a stacking gel (125 mM Tris-HCl pH 6.8, 0.1 % (w/v) SDS, 0.05 % (w/v) ammonium persulphate (APS), 4.0 % (v/v) acrylamide, 0.1 % (v/v) bisacrylamide) overlaying the resolving gel.

Up to 20 µl of denatured protein sample was loaded in each well. Molecular weights of sample proteins were determined by loading 10 µl of a molecular weight marker per gel (Precision Plus Protein™ Standards, Bio-Rad). The Precision Plus Protein Standard contained ten protein bands of 10 kD, 15 kD, 20 kD, 25 kD, 37 kD, 50 kD, 75 kD, 100 kD, 150 kD, and 250 kD.

Proteins were separated by electrophoresis in usually, a 12 % (v/v) acrylamide gel (375 mM Tris-HCl pH 8.8, 12.0 % (v/v) acrylamide, 0.3 % (v/v) bisacrylamide, 0.15 % APS, 0.1 % (w/v) SDS, 0.08 % (v/v) TEMED, as described above unless otherwise stated) in 1 × electrophoresis running buffer (section 2.1.1.3). Samples were run at 200 volts for approximately 45 minutes.

2.4.2 Coomassie Staining

Protein samples resolved by SDS-PAGE were stained with Coomassie blue stain (section 2.1.1.3) for 1 hour. Excess stain was removed by washing the gel for 20 minutes, two to three times with de-staining solution (section 2.1.1.3) with gentle agitation. Gels were then placed in water containing glycerol for 30 minutes and dried using a Bio-Rad drying system according to manufacturers instructions.

2.4.3 Western blotting

Following separation of the protein samples by SDS-PAGE, the proteins were transferred to a nitrocellulose membrane (BioRad) ready for detection with a specific primary antibody. Protein blotting was performed using a semi dry transfer cell (Bio-Rad). The gel was placed on a piece of nitrocellulose membrane (0.45 μm thick, Bio-Rad) and sandwiched between blotting paper soaked in 1 × transfer buffer (section 2.1.1.3). The entire sandwich was then placed within the semi dry transfer cell according to manufacturers instructions. Proteins were transferred from the gel to the nitrocellulose membrane at 0.8 Amps for 20 minutes at 20 volts.

After transfer of the resolved protein samples onto the nitrocellulose membrane, non-specific binding was abolished by blocking the membrane with blocking buffer (section 2.1.1.3) at 4°C overnight.

The membrane was then incubated in blocking buffer containing primary antibodies (table 2.4, A) at specific concentrations (as described in the relevant sections) for 1 hour at approximately 22 °C on a shaking platform. The nitrocellulose membrane was washed 5 times in 1 × PBS containing 0.05 % (v/v) Tween-20 (pH 7.4). Secondary antibody was then diluted (see table 2.4, B for working dilutions of secondary antibodies) in fresh blocking buffer, added to the membrane, and incubated at room temperature for 1 hour on a shaking platform. Excess, unbound secondary antibody was removed by washing the membrane 5 times in 1 × PBS containing 0.05 % (v/v) Tween-20 (pH 7.4).

Secondary antibody binding was detected using the ECL Plus Western Blotting Detection Reagents (Amersham Biosciences) in accordance with the manufacturers instructions. The ECL Plus Western Blotting Detection system is based on the combined HRP and peroxide catalyzed oxidation of a Lumigen PS-3 Acridan substrate. The resulting light was detected using Fuji films (Jet X-ray, UK).

(A)

Name	*Mono/Polyclonal*	*Specificity*	*Affinity purified*	*Raised in*	*Supplier*	*Western blot Titre*	*Immunocytochemistry titre*
SG1703	Polyclonal (C-terminus peptide)	NHS peptide sequence	No	Rabbit	Sigma Genosys	1:1000	1:1000
Myc	Monoclonal	Myc peptide sequence	Yes	Mouse	Sigma-Aldrich	1:1000	1:2000
p34	Polyclonal	p34 subunit of the Arp2/3 complex	Yes	Rabbit	-	-	1:50

(B)

Specificity	*Affinity purified*	*Raised in*	*Supplier*	*Conjugate*	*Western blot Titre*	*Immunocytochemistry titre*
Goat IgG (H+L)	Yes	Rabbit	Jackson ImmunoResearch	Peroxidase-conjugated	1:30,000	-
Rabbit IgG (H+L)	Yes	Donkey	Jackson ImmunoResearch	Cyanine 3-congugated	-	1:100
Mouse IgG (H+L)	Yes	Donkey	Jackson ImmunoResearch	Cyanine 2-congugated	-	1:100

Table 2.4 (A) List of primary antibodies used and, (B) secondary antibodies used to detect primary antibodies

2.4.4 Expression and Purification of Glutathione-S-Transferase (GST) Fusion Proteins

2.4.4.1 Initial analysis of GST-fusion proteins

A 5 ml aliquot of LB broth containing ampicillin was inoculated with the relevant pGEX-2T clone and grown overnight at 37°C with vigorous shaking (transformant culture). The transformant culture was then diluted 1 in 100 into 10 ml fresh LB broth containing ampicillin, and incubated with vigorous shaking at 37°C for 2 hours. 1 ml aliquot from each liquid culture was removed and placed on ice for use as non-induced control samples. To the remaining 9 ml of culture, IPTG at a final concentration of 1 mM was added, and the cultures incubated with vigorous shaking at 37°C for 3 hours. Following incubation, 40 µl of each of the induced and uninduced cultures were transferred to a sterile 1.5 ml eppendorf tube, to which 10 µl of 5 × SDS sample buffer was added. Samples were then denatured by heating at 95°C for 5 minutes before loading onto a 12 % SDS-PAGE gel (see section 2.4.1). The remaining induced cultures, were centrifuged at 4,000 ×g for 10 minutes to pellet the cells. The supernatant was then aspirated and the pellets stored at – 20 °C for future use.

2.4.4.2 Purification of GST fusion proteins

Milligram quantities of GST fusion protein were purified following initial confirmation that the fusion protein was being expressed (see section 2.4.4.1).

Starter cultures (section 2.4.4.1) were diluted 1:100 into 100 ml LB/ampicillin and grown at 37 °C with vigorous shaking for 12 to 15 hours. The culture was then diluted 1:10 into 1 litre fresh LB/ampicillin medium, which was split between two 2-litre flasks, and grown at 37 °C with vigorous shaking for 1 hour. A 5 ml volume was then removed and transferred to a clean sterile universal tube. To the remaining culture, 100 mM IPTG was added to a final concentration of 0.1 mM and the induced culture grown for an additional 3 hours. The 5 ml volume taken out prior to the addition of IPTG (the uninduced control sample), was also grown for an additional 3 hours. The induced culture was then centrifuged for 10 minutes at 5,000 × g, at approximately 22°C to pellet the cells. The supernatant was discarded and the pellet resuspended in 20 ml 1 × PBS containing protease inhibitor cocktail as mentioned previously (section 2.1.1.3).

Resuspended cells were then lysed on ice using a 5 mm diameter probe sonicator. 10 % (v/v) Triton X-100 was then added to the lysed cells to 1 % final concentration and mixed. Lysed cells were then centrifuged at 10,000 × g for 5 minutes at 4 °C, and the supernatants collected. The supernatant was transferred to a tube containing 1.5 ml 50 % slurry of glutathione-agarose beads (Amersham Biosciences) and mixed by end-over-end rotation for 5 minutes at approximately 22 °C.

The slurry was then washed by adding up to 50 ml ice-cold 1 × PBS, mixed, and centrifuged for 10 seconds at 500 × g, approximately 22 °C. PBS was aspirated and the wash repeated an additional three times. The beads were resuspended in 1 ml ice-cold 1 × PBS and transferred to a clean sterile 1.5 ml microcentrifuge tube. The sample was centrifuged at 500 × g, approximately 22 °C, to pellet the beads. The supernatant was then aspirated and the fusion protein eluted by the addition of 1 ml 50 mM Tris-HCl (pH 8.0)/10 mM reduced glutathione. The sample was mixed gently for 2 minutes, centrifuged for 10 seconds at 500 × g, approximately 22 °C, and the supernatant collected. The elution step was repeated twice. Samples were analysed by SDS-PAGE (section 2.4.1) and stored long term at – 80 °C in 10 % glycerol.

2.5 Cell Based Techniques

All cell culture lines were grown in an atmosphere of 5 % CO_2 at 37 °C and were passaged every two to three days, or when approximately 90 % confluent.

2.5.1 Cell Culture Maintenance

2.5.1.1 Adherent cells

Table 2.5 lists the media used to maintain all the mammalian cell lines used in this work. Except for MTLn3 cells, all cells were cultured in the presence of 10 % (v/v) foetal bovine serum and 50 μg/ml gentamycin. MTLn3 cells were cultured in 5 % (v/v) foetal bovine serum 50 μg/ml gentamycin.

Cell type	*Abbreviation*	*Morphology*	*Media*
Human Caucasian colon adenocarcinoma	Caco-2	Epithelial	DMEM (1×) liquid + L-Glutamine, 4500 mg/L D-Glucose, 110 mg/L Sodium Pyruvate
Chinese hamster ovary	CHO	-	DMEM/F12 (1×) liquid + GlutaMAX I
Monkey African Green Kidney	COS-7	Fibroblast	DMEM/F12 (1×) liquid + GlutaMAX I
Dog cocker spaniel kidney	MDCK	Epithelial	DMEM (1×) liquid + L-Glutamine, 4500 mg/L D-Glucose, 110 mg/L Sodium Pyruvate
Metastatic rat mammary adenocarcinoma	MTLn3	Epithelial	MEM-α (1×) liquid + Earl salts and L-Glutamine. No ribonucleosides or Deoxyribonucleosides

Table 2.5 Adherent cell lines, their morphology, and media used.

2.5.2 Storage of cells

Cells to be stored in liquid nitrogen were pelleted, resuspended in 2 mls cell culture freezing media (Gibco, UK) and placed in 100 % (v/v) ethanol in a cell freezing container at – 80 °C for 24 hours before freezing in liquid nitrogen.

2.5.3 Reviving Stored Cells

Cell aliquots frozen in liquid nitrogen were removed and thawed quickly in a water bath at 37 °C. The thawed cells were then immediately transferred to pre warm media and left at 37 °C in an atmosphere of 5 % CO_2.

2.5.4 Determination of total cell counts and viable cell number

Viable cells in culture were counted using Trypan Blue. Trypan Blue only crosses the membrane of dead (non-viable) cells therefore appearing dark blue. Adherent cells were treated with Trypsin (Gibco, UK) and resuspended in 4 ml medium. A 50 µl aliquot was transferred to a separate tube containing 50 µl of 0.4 % (w/v) Trypan Blue solution, mixed and left to stand for 5 minutes. A cover slip was placed on a clean hemocytometer and the Trypan Blue-cell suspension transferred to each chamber by

capillary action. Viable and non-viable cells were counted in the 1 mm centre square of chamber 1 followed by the four 1 mm corner squares, and was repeated for chamber 2. Non-viable cells were then counted as a percentage of viable cells.

2.5.5 LipoFectamine Plus

LipoFectamine Plus, a cationic lipid, was used to transfect mammalian cells with plasmid DNA. In order to achieve maximum transfection efficiency, the following conditions were optimised: cell number, plasmid DNA and lipofectamine concentrations, and the length of exposure of cells to DNA-liposome complexes.

2.5.5.1 Transfection of CHO cells using LipoFectamine Plus

CHO cells were seeded at a density of approximately 30,000 cells per well on an eight well chamber slide, and left to grow and adhere for one to two days before transfecting. Once the cells were approximately 80 % confluent, they were transfected with the plasmid DNA construct of interest. Two solutions (A and B) were prepared. Solution A consisted of 100 ng of plasmid DNA, 1 µl of Plus reagent, and 25 µl of serum free DMEM/F12 media. Solution B contained 0.5 µl Lipofectamine and 100 µl of serum free DMEM/F12 media. Both solutions were mixed separately and left at 22 °C for fifteen minutes. The two solutions were then mixed together and left at 22 °C for an additional fifteen minutes to allow the DNA-liposome complexes to form. The CHO/MTLn3 cells to be transfected were then washed twice in Hanks' Balanced Salt Solution (HBSS), and a 100 µl of the DNA-liposome solution added to each well of the eight well chamber slide. The cells were then incubated for 3 - 4 hours at 37 °C in an atmosphere of 5 % CO_2. Following the incubation period, an equal volume of DMEM/F12 containing 20 % (v/v) fetal calf serum (10 % (v/v) final concentration), and 2 % (v/v) streptomycin/penicillin (1 % (v/v) final concentration), was added to each well of the eight well chamber slide and the cells incubated for 24 hours at 37 °C with 5 % CO_2. Transfected cells were then stained and analysed (see section 2.5.6.1).

2.5.6 Immunocytochemistry

2.5.6.1 Adherent cells

Glass chamber slides and cover slips were pre-treated prior to seeding MTLn3 cells by first washing in 1M HCl for 3-5 minutes followed by two washes with sterile 1× PBS. PBS was then aspirated away and 95 % (v/v) ethanol overlaid on the chamber slide/cover slip for 1-2 minutes. A second wash with 1× PBS was then carried out before seeding the cells. No pre-treatment was necessary for other cell types used in this study.

Transfected CHO cells seeded on 8 well chamber slides at a density of approximately 30,000 and seeded MTLn3 cells, were rinsed twice in 1 × PBS and fixed in 4 % (w/v) paraformaldehyde in 1 × PBS for fifteen minutes at 22 °C. The cells were then washed twice in 1 × PBS and permeabilised by incubating the cells at 22 °C in 0.1 % (v/v) Triton X 100 in 1 × PBS for fifteen minutes. The cells were then washed twice in 1 × PBS for five minutes. To prevent non-specific binding, the cells were incubated in blocking buffer (1 % (w/v) bovine serum albumin (BSA), 10 % (v/v) donkey serum in 1 × PBS) for 45 minutes at 22 °C. Seeded MTLn3 cells were incubated in blocking buffer for 45 minutes at 22 °C in the presence of either rhodamine-, Alexa Fluor 488-, or Alexa Fluor 647 phalloidin (see table 2.6) at a final concentration of 20 units/ml. The blocking buffer was then replaced with fresh blocking buffer containing the appropriate dilution of primary antibody (see table 2.4, A) and left at 22 °C for 1 hour. The cells were then washed five times with 1 × PBS. A 100 µl aliquot of fresh blocking buffer containing a 1:100 dilution of the appropriate secondary antibody conjugated with either cyanine-2 or –3 (see table 2.4, B) was added to each well and incubated at approximately 22 °C for 1 hour. The cells were then washed five times in 1 × PBS. On the penultimate wash, the 1 × PBS contained 4'-6 Diamidino-2-phenylindole (DAPI) at a concentration of 0.002 mg/ml. The final wash in 1 × PBS was aspirated and one drop of binding media (90 % glycerol in 1 × PBS) added to each well before overlaying a cover slip.

Phallotoxin probe	Conjugate	Stock concentration (units/ml)	Working concentration (units/ml)
Alexa Fluor 488 phalloidin	Green-fluorescent Alexa Fluor 488	200	20
Alexa Fluor 647 phalloidin	Far-red fluorescent Alexa Fluor 647	200	20
Rhodamine phalloidin	Tetramethylrhodamine (TRITC)	200	20

Table 2.6 List of conjugated phallotoxins used to stain the actin cytoskeleton of MTLn3 cells.

2.5.7 Epidermal growth factor (EGF) stimulation of MTLn3 cells

MTLn3 cells were seeded onto small glass cover slips placed within a 35 mm dish at a density of approximately 240,000 cells. Glass cover slips were pre-treated as stated in section 2.5.6. Seeded cells were left for 24 hours or until approximately 80 % confluent. Cells were then starved for a maximum of 3 hours by replacing the media for serum-free MEM-alpha media supplemented with 0.35 % (w/v) BSA. An equal volume of serum-free MEM alpha media containing 10 nM EGF solution was then added to the starved cells and left for 3 minutes to allow for formation of lamellipods. After 3 minutes the media was removed and 4 % paraformaldehyde added to fix the cells. The cells were then fixed and stained accordingly (see section 2.5.6). Upon staining of the stimulated MTLn3 cells, a drop of binding media was added to each of the cover slips before being inverted and mounted onto a glass slide.

2.6 Actin polymerisation assay

Measuring polymerisation through fluorescence enhancement of pyrene-conjugated actin is a sensitive, versatile and quick assay to help determine if a protein activates the Arp2/3 complex. Control and test samples are mixed together with a small amount of pyrene-conjugated actin and the Arp2/3 complex and fluorescence measured using a spectrophotometer. Fluorescence is enhanced up to twenty fold by the association of actin monomer into the polymer form.

Bovine brain Arp2/3 complex and rabbit skeletal muscle actin (> 95 % purity), were purchased from Cytoskeleton Inc along with an Actin Polymerisation Biochem Kit. The Actin Polymerisation Biochem Kit contained: actin (> 99 % purity) modified to

contain co-valently linked pyrene at cysteine 374, General Actin Buffer, Actin Polymerisation Buffer, 100 mM ATP pH 7.0, and 100 mM Tris-HCl pH 7.5.

To 1 ml General Actin Buffer, 2 µl of ATP solution was added and mixed (A buffer). 225 µl and 605 µl aliquots of A buffer were then mixed with 5 µl pyrene-labelled actin and 20 µl actin respectively, and left at 22 °C for 1 hour. The actin monomer stocks were then centrifuged at 150,000 ×g for 2 hours at 4°C and the top 80 % of supernatant aspirated for further use in the polymerisation assays. This step reduces the endogenous nucleation centres formed from freeze/thawing the actin samples and reduces the starting background level of fluorescence.

A control assay lacking the test substance was performed for each run. To a clean quartz cuvette (3 × 3 mm light path), 4.5 µl pyrene-labelled actin and 87.75 µl actin were added (5 % pyrene labelled actin). A 0.5 µl aliquot of Arp2/3 complex (13 nM final concentration) was then added to the cuvette followed by 9 µl of Actin Polymerisation Buffer, mixed thoroughly, and fluorescence measured immediately using a Perkin Elmer LS 50B luminescence spectrometer. Fluorescence was measured for a period of 10 minutes or until a plateau in signal strength was achieved. Excitation was set at a wavelength of 365 nm, with a bandwidth of 15 nm and the emission wavelength at 407 nm, with a bandwidth of 10 nm.

GST-tagged NHS constructs (in 10mM Tris pH 8.0, 50 mM NaCl; section 2.4.4) at varying concentrations, were then tested for their ability to compete with the WASP WCA domain for binding to the Arp2/3 complex. A competitive assay consisted of the above assay being performed in the presence of both GST-tagged WASP WCA (100 nM, in 20 mM Tris pH 7.5, 25 mM KCl, 1 mM $MgCl_2$, 0.5 mM EDTA, 0.2 % (w/v) dextran, 2 % (w/v) sucrose buffer, Cytoskeleton Inc.) and the GST-tagged NHS construct. GST-tagged WASP WCA was added after the addition of the GST-tagged NHS construct due to the WASP WCA domain being a potent activator of the Arp2/3 complex.

2.7 Bioinformatics and computational software

A series of bioinformatic tools were used to define the genomic interval Xp22.13 - 22.2, identify polymorphic markers and characterise genes within the region.
The main sites used for this study were as follows:

- UCSC (http://genome.ucsc.edu/),
- NIX and PIX (http://www.hgmp.mrc.ac.uk/),
- UniGene (NCBI, http://www.ncbi.nlm.nih.gov/),
- Genecards (http://bioinformatics.weizmann.ac.il/cards/),
- Ensembl (http://www.ensembl.org/),
- Sanger Institute (http://www.sanger.ac.uk/),
- ExPASY (http://ca.expasy.org/).

Novel microsatellites residing within the CXN interval were sought using an in house nucleotide repeat locator programme (World-wide Webster web designs 2000).

2.7.1 UCSC genome browser

The UCSC genome browser allows users to visualise a number of genomes including human, mouse and rat. The browser displays assembly contigs and gaps, mRNA and EST alignments, multiple gene predictions, homologies to other species, single nucleotide polymorphisms, sequence-tagged sites and more as a series of tracks aligned with the genomic sequence of interest (Kent *et al.*, 2002; Karolchik *et al.*, 2003).

2.7.2 Nucleotide Identification of unknown sequences (NIX)

NIX, a WWW tool, runs numerous DNA programmes on a denoted DNA sequence allowing for a general consensus to be seen between the programmes. Such programmes include exon prediction programmes (GRAIL, Fex, Hexon, MZEF), gene prediction programmes (Genemark, Genefinder, FGene), BLAST, Polyah (polyadenylation-signal identifier), RepeatMasker and tRNAscan (transfer RNA identifier).

2.7.3 Protein Identification of unknown sequences (PIX)

PIX allows users to view results of numerous peptide analysis programme's run simultaneously on a peptide sequence of interest. The type of analysis runs on a query peptide sequence include predicting, secondary structure (e.g. Predator); coiled coil (coil), transmembrane (DAS), signal peptide (Signal) and helix-turn-helix (HTH) domains; and running Blast searches against sequence and the (SPTR, NRL_3D) domain (SBASE, PRODOM) databases.

2.7.4 ExPASy tools

The ExPASy (Expert Protein Analysis System) proteomics server (Swiss Institute of Bioinformatics, SIB), provides databases and analysis programmes for protein sequences and structures (Gasteiger *et al.*, 2003). A software programme used in this thesis included T-coffee, an alignment algorithm allowing multiple alignment of protein sequences. The PROSITE database (a database of protein families and domains) was used to identify potential protein domains and homologies to other proteins or classes of protein. Identified homologies to proteins or specific domains were then searched for in the Swiss-Prot protein knowledge database to elucidate their function and highlight conserved/functionally important residues.

2.7.5 DNAStar

Nucleic acid and protein sequences were analysed using DNAStar computational software (DNAStar, Inc., USA). Sequencing data obtained from clones or PCR products were analysed using SeqMan. EditSeq was used to analyse downloaded Genbank sequences. Pairwise and multiple sequence alignments were performed using MegAlign, whilst MapDraw was used to generate a map of restriction endonuclease sites present within a given nucleic acid sequence.

- Chapter Three -

Refinement of the X-linked cataract locus (CXN) and gene analysis for CXN and Nance-Horan syndrome (NHS)

3.1 Introduction

The chromosomal region Xp22 harbours numerous eye disease genes including Nance-Horan Syndrome (NHS; nuclear cataract, microphthalmia and microcornea, MIM 302350), X-linked congenital cataract (CXN; Francis *et al.*, 2002), retinitis pigmentosa (RP23; Hardcastle *et al.*, 2000, MIM 300424), and retinoschisis 1 (RS1, MIM 312700). The causative genes, for NHS, CXN, and RP23 are not yet known and remain to be identified. This research project focuses on two diseases which mapped to this genomic region, Nance-Horan Syndrome and X-linked congenital cataract.

3.1.1 X-linked congenital cataract (CXN)

An X-linked congenital cataract (CXN) locus was recently mapped within a five-generation family to a 3 cM interval on chromosome Xp22.13 – 22.2 flanked by markers DXS9902 and DXS999 (Zmax = 3.64 at $\theta = 0$ for marker DXS8036; figure 3.1; Francis *et al.*, 2002).

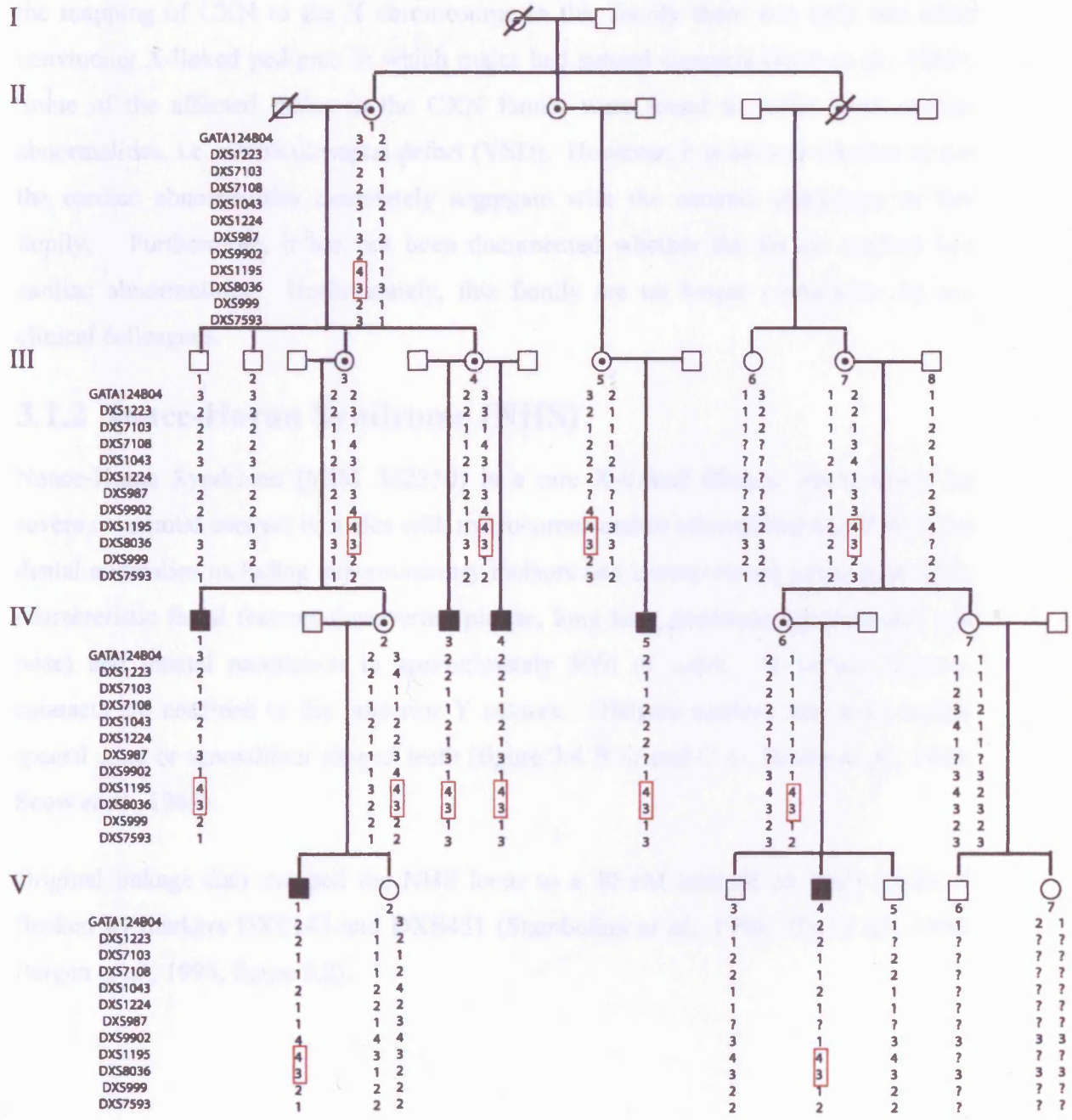


Figure 3.1 CXN pedigree showing haplotypes on Xp22.13 – 22.2

Boxed haplotypes of affected males and carrier females show recombination with markers DXS99902 and DXS999, as drawn in original paper (Francis *et al.*, 2002). Disease segregates with markers DXS1195 and DXS8036.

Within this five-generation family, males are severely affected with a total opacity of the lens and require cataract surgery within the first few months of life. Females have a milder form of disease consistent with an X-linked carrier status. The cataract phenotype in carrier females is one of sea-fan of nuclear opacity (figure 3.4A). Until the mapping of CXN to the X chromosome in this family there was only one other convincing X-linked pedigree in which males had sutural cataracts (Krill *et al.*, 1969). Some of the affected males in the CXN family were found to suffer from cardiac abnormalities, i.e. ventriculoseptal defect (VSD). However, it is unclear whether or not the cardiac abnormalities completely segregate with the cataract phenotype in this family. Furthermore, it has not been documented whether the female carriers had cardiac abnormalities. Unfortunately, this family are no longer contactable by our clinical colleagues.

3.1.2 Nance-Horan Syndrome (NHS)

Nance-Horan Syndrome (MIM 302350) is a rare X-linked disease characterised by severe congenital cataract in males with microcornea and/or microphthalmia, distinctive dental anomalies including supernumerary incisors and crown shaped permanent teeth, characteristic facial features (anteverted pinnae, long face, prominent nasal bridge and nose) and mental retardation in approximately 30% of cases. In carrier females, cataracts are confined to the posterior Y sutures. Obligate carriers also have widely spaced cone or screwdriver shaped teeth (figure 3.4 B iii and C iii; Bixler *et al.*, 1984; Seow *et al.*, 1985).

Original linkage data mapped the NHS locus to a 30 cM interval on Xp22.13-22.31 flanked by markers DXS143 and DXS451 (Stambolian *et al.*, 1990; Zhu *et al.*, 1990; Bergen *et al.*, 1994, figure 3.2).

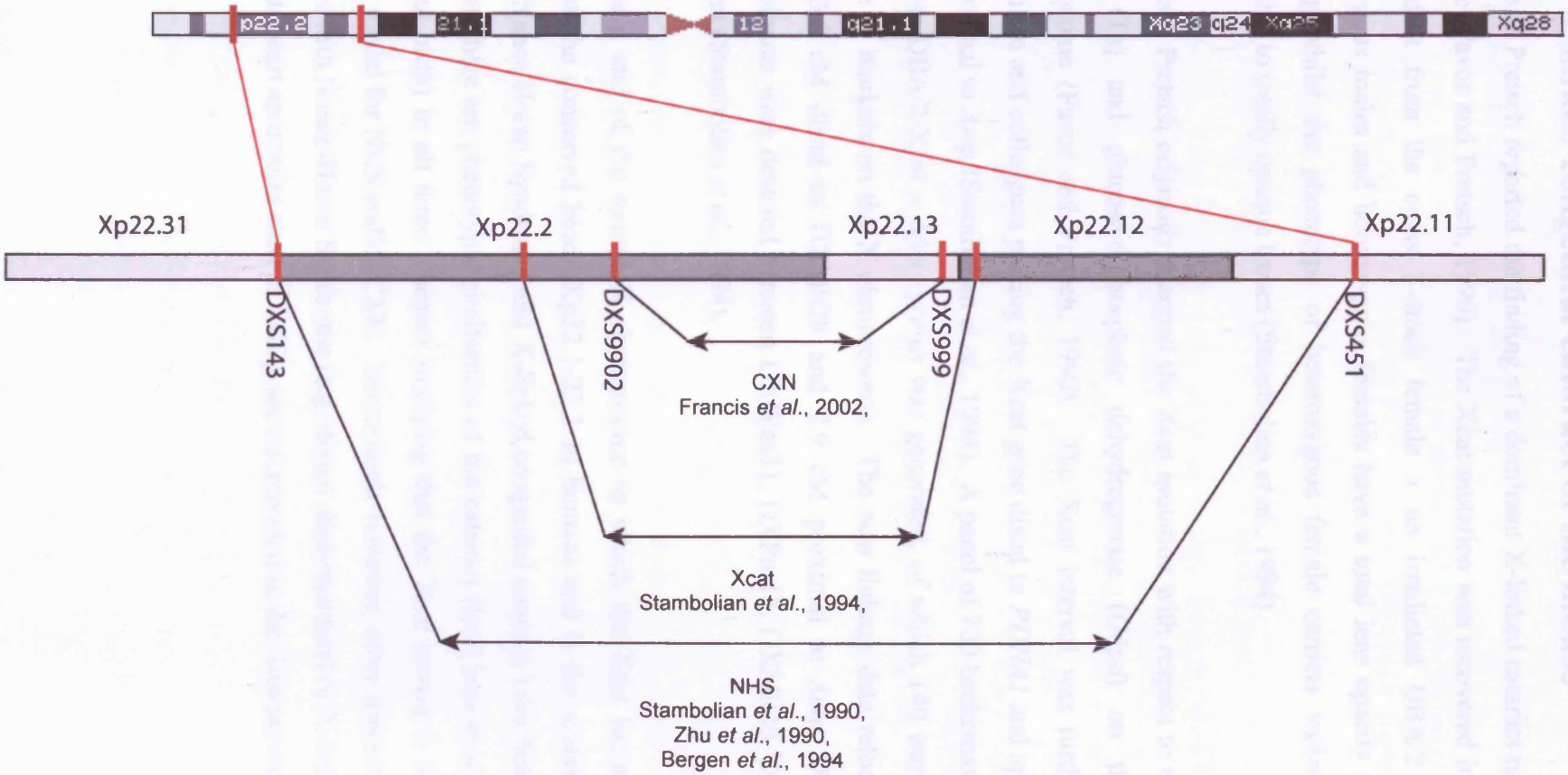


Figure 3.2 Schematic diagram of the genomic intervals for CXN, NHS, and Xcat.

Note, the Xcat genomic interval mapping to the mouse chromosomal region syntenic with the Xp22.1-22.3 block, is superimposed on the human X chromosome for comparison.

3.1.3 X-linked congenital cataract of the mouse

Favor and Pretsch reported the finding of a dominant X-linked cataract mutation (Xcat) in mice (Favor and Pretsch, 1990). The Xcat mutation was recovered in an F_1 female descendent from the cross T-stock female × an irradiated DBA/2 male. Both hemizygous males and homozygous females have a total lens opacity present at eye opening whilst the phenotype of heterozygous female carriers varies from barely noticeable to totally opaque lenses (Stambolian *et al.*, 1994).

Favor and Pretsch originally mapped the Xcat mutation with respect to two mutations, tabby (Ta) and glucose-6-phosphate dehydrogenase (*G6pd*) on the mouse X chromosome (Favor and Pretsch, 1990). The Xcat interval was further refined by Stambolian and colleagues placing the Xcat gene distal to *PDHA1* and approximately 3 cM proximal to *Amg* (Stambolian *et al.*, 1994). A panel of 700 backcross progeny from the cross DBA/2-Xcat × *Mus spretus* was generated, of which 140 were analysed for linkage to markers on the X chromosome. The new linkage data relocated the Xcat locus 3.5 cM distal to DXMit28 and 2.9 cM proximal to *Amg*, DXMit12. No recombinants were detected between DXWas31, DXPas18, DXMit20, DXMit15, *Grpr* and Xcat (Stambolian *et al.*, 1994).

The distal end of the mouse X chromosome to which the Xcat locus was mapped contains the conserved block Xp22.1-22.3 in humans and is the syntenic interval to which Nance-Horan Syndrome and X-linked congenital cataract have been mapped. In addition, there are phenotypic similarities of the cataract (total lens opacity in affected males at birth) in all three diseases implying that the Xcat mouse is likely to be an animal model for NHS and/or CXN. Interestingly however, other abnormalities seen in patients with Nance-Horan Syndrome (e.g. dental abnormalities) or X-linked congenital cataract (heart anomalies such as VSD), are not reported in the Xcat mouse.

3.2 Results

3.2.1 Families and phenotypes

The CXN family previously mentioned in section 3.1.1 was included in this study, along with four families with a diagnosis of Nance-Horan syndrome (figure 3.3).

3.2.1.1 NHS family 1

A four generation family diagnosed with Nance-Horan Syndrome (NHS) was evaluated (figure 3.3, B). Two brothers were noted soon after birth to have bilateral dense congenital cataracts and microphthalmia (IV:2 and IV:3, figure 3.3, B). The youngest brother (IV:3) also had bilateral cleft palate, which was repaired in infancy. At a later age both brothers were found to have abnormal teeth and typical dysmorphic features (anteverted pinnae, long face, prominent nasal bridge and nose) consistent with a diagnosis of Nance-Horan syndrome (NHS). Their mother (III:3) had bilateral sutural lens opacities consistent with being an X-linked carrier of the disorder.

3.2.1.2 NHS family 2

Family two (figure 3.3, C) exhibited typical features of NHS. One affected son (III:2) exhibited dense central nuclear opacities along with large anteverted pinnae, typical NHS dysmorphism (long face, prominent nasal bridge and nose), diastema and prominent notching of incisors (figure 3.4, C iii-iv). The other affected son (III:1) displayed dense white central lens opacity with denser central star shaped opacity, but no typical NHS features were documented. The affected sons mother (II:1) had posterior polar and cortical cataract (figure 3.4, C i-ii), and it was reported she had poor dental enamel.

3.2.1.3 NHS family 3

The affected male (II:1) of family 3 (figure 3.3, D) displayed typical features of NHS (prominent teeth, anteverted pinnae, long face, and broad nasal bridge) in addition to severe mental retardation, epilepsy and hypotonia (figure 3.4, B i). The affected sons mother (I:1) had sutural cataract and displayed screwdriver shaped teeth (figure 3.4, B ii and iii).

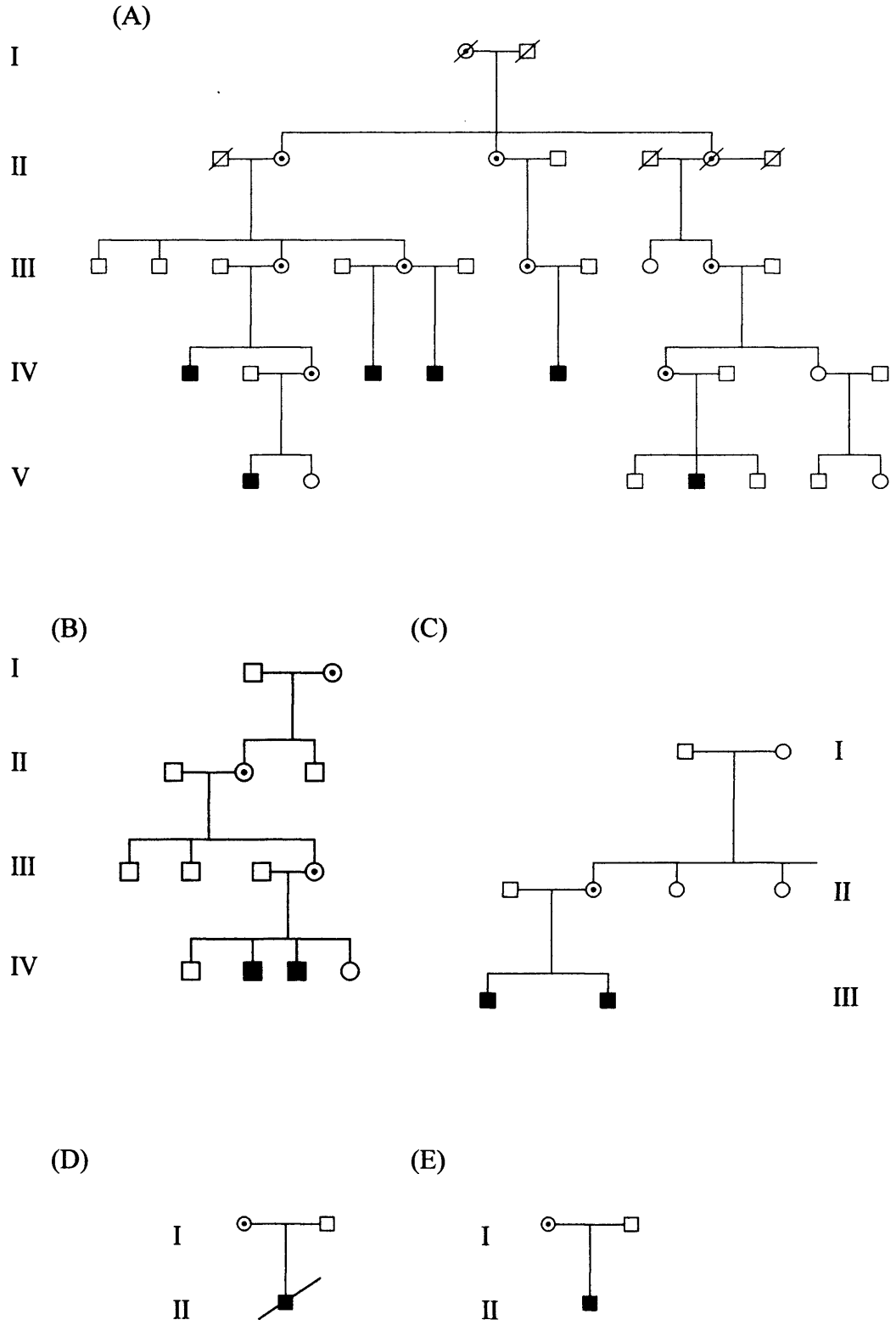


Figure 3.3 Pedigrees of X-linked congenital cataract and Nance-Horan Syndrome
CXN pedigree (A), NHS 1 pedigree (B), NHS 2 pedigree (C), NHS 3 pedigree (D), NHS 4 pedigree (E). Squares denote males and circles, females. Blackened squares are indicative of an affected individual and circles with black dots represent carrier females. Symbols with a diagonal line through them represent deceased individuals.

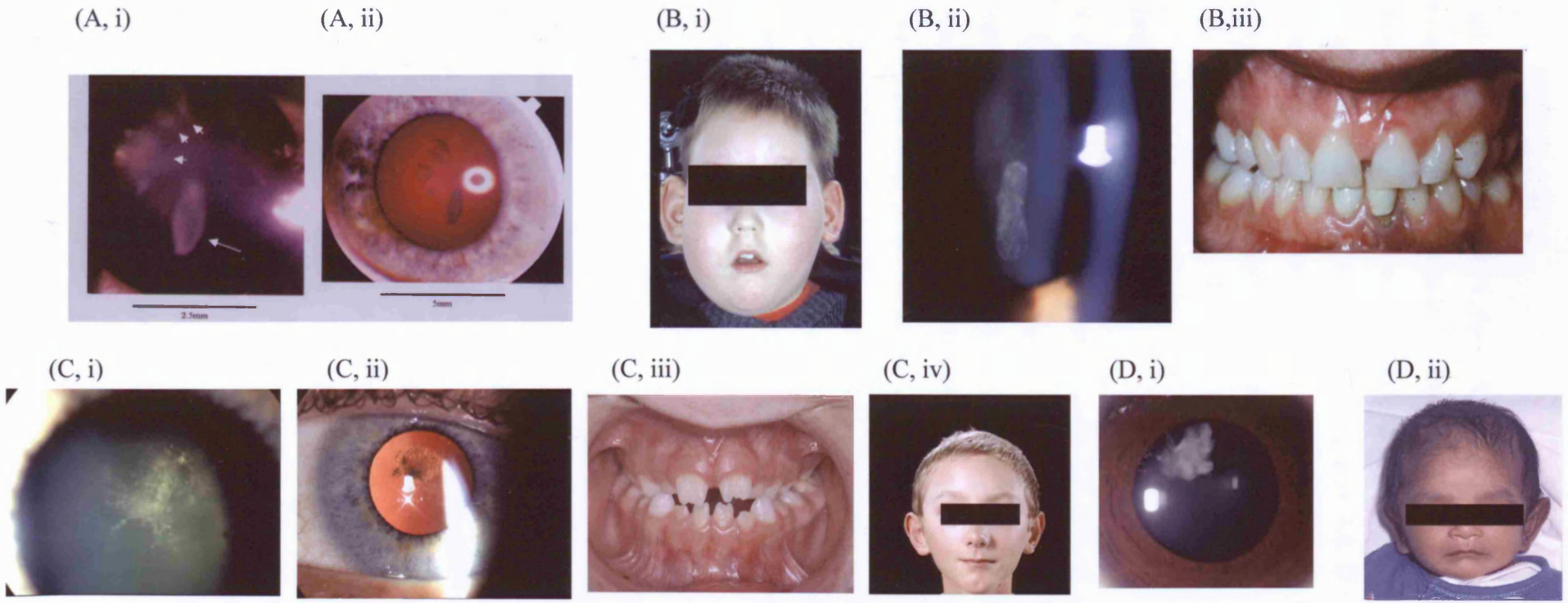


Figure 3.4 Specific clinical features of CXN and NHS.

Slit lamp view of a heterozygous 50 year old female from the CXN family, showing fan of nuclear lens opacification (A, i). Same eye at a lower magnification and with the lens in retroillumination (A, ii). An Affected male (II:1) with typical NHS features and severe developmental delay from NHS family 3 (B, i). Sutural cataract (B, ii) and screwdriver shaped teeth (B, iii) in a female carrier for NHS (I:1, NHS family 3). Slit lamp view (C, i) and lens retroillumination (C, ii) of a 34 year old carrier female (II:1, NHS family 2). Teeth anomalies in an affected male of NHS family 2 (III:2, C, iii). An Affected male (III:2) with typical NHS features from NHS family 2 (C, iv). Slit lamp view (D, i) of an affected mother in NHS family 4 (I:1). An Affected male (II:1) with typical NHS features from NHS family 4 (D, ii).

3.2.1.4 NHS family 4

The affected male in this family (II:1, figure 3.3, E) had dense white central lens opacity along with an upside down Y sutural opacity with a milky lens opacity diffuse edge. The periphery of the lens was clear. The affected sons mother (I:1) exhibited fine dots at both posterior Y sutures and cortical wedges (figure 3.4, D i). Other typical NHS features were not documented, although the family do have a history of abnormally large ears as seen in the affected son (II:1, figure 3.4, D ii).

3.2.2 Genotyping

The CXN disease interval spans approximately 3.5 Mb of genomic sequence on Xp22.13 – 22.2 (Francis *et al.*, 2002, figure 3.2). Prior to further genetic and genomic studies, segregation at Xp22 was confirmed through genotyping a total 46 known microsatellite markers (table 3.1) spanning the X chromosome, in 20 individuals of the CXN family. Genotyping of the 46 known microsatellite markers was outsourced to The John Vane Science Centre (Barts, London), the results were scored twice by two individuals and then compared with our own, previous data. The results (figure 3.5) confirmed disease segregation in the CXN family to chromosomal region Xp22.13 – 22.2, with markers DXS1053-DXS8036-DXS1195 (figure 3.5).

Since the disease interval in the CXN family on Xp22.13-22.2 spans approximately 3.5 Mb of DNA and the average gene density in humans is 1 gene per 100,000 bases, there was the potential for the CXN interval to harbour up to approximately 35 genes. In an attempt to refine the CXN interval and therefore reduce the number of potential candidate genes for analysis, additional and novel microsatellite markers were sought for genotyping in the CXN family. Microsatellite markers DXS1053, DXS1195, and AFMB343YD5, within the CXN disease interval were selected and primer sequences obtained from the genome database (http://www.gdb.org, table 3.2). In addition, a further 6 novel di-nucleotide repeats (S1, S2, S3, S4, S6, and S7), two tri-nucleotide repeat markers (S5 and S9), and one tetra-nucleotide repeat marker (S8), were identified from available genomic sequence within the disease interval using an in-house repeat locator programme (World-wide Webster web designs 2000).

Marker name	Chromosome location	Heterozygosity	Genetic distances (Mb) marker to marker
DXS1060	Xp22.32	-	
DXS1223	Xp22.31	0.75	2.83
DXS7108	Xp22.22	0.75	1.50
DXS9902	Xp22.2	-	4.88
DXS1053	Xp22.2	0.84	0.04
DXS8036	Xp22.2	0.74	1.28
DXS1195	Xp22.13	0.61	0.29
DXS999	Xp22.13	0.75	1.10
DXS7593	Xp22.11	0.70	3.36
DXS1061	Xp21.3	0.73	4.85
DXS1214	Xp21.2	0.79	3.62
DXS8090	Xp21.1	0.81	5.49
DXS1068	Xp11.4	0.82	1.75
DXS8102	Xp11.4	0.65	0
DXS8015	Xp11.4	0.74	0.50
DXS993	Xp11.4	0.79	1.16
DXS8080	Xp11.3	0.68	2.90
DXS8083	Xp11.3	0.74	0.80
DXS1055	Xp11.3	0.72	0.98
DXS1039	Xp11.23	0.61	2.78
DXS991	Xp11.21	0.82	6.04
DXS1216	Xq13.1	0.61	12.54
DXS986	Xq21.1	0.76	10.84
DXS1196	Xq21.31	0.79	7.11
DXS1217	Xq21.31	0.62	1.51
DXS8077	Xq21.33	0.60	6.68
DXS8020	Xq22.1	0.78	4.10
DXS1106	Xq22.2	0.67	2.96
DXS1059	Xq23	0.72	8.39
DXS8088	Xq23	0.62	1.83
DXS8055	Xq23	0.64	1.11
DXS8064	Xq24	0.56	2.38
DXS8067	Xq24	0.70	1.89
DXS1001	Xq24	0.82	0.28
DXS8009	Xq25	0.52	6.10
DXS1047	Xq26.1	0.81	2.70
DXS1062	Xq26.3	0.74	8.00
DXS984	Xq27.1	0.72	2.13
DXS1205	Xq27.1	0.72	0.43
DXS1227	Xq27.2	0.73	0.34
DXS8106	Xq27.3	0.69	1.18
DXS8043	Xq27.3	0.81	1.62
DXS8045	Xq27.3	0.49	1.29
DXS998	Xq27.3	0.57	0.90
DXS8069	Xq28	0.62	2.79
DXS1073	Xq28	-	3.88

Table 3.1 Micosatellite markers used for genetic characterisation of the CXN family
A total of 46 microsatellite markers spanning the X chromosome were genotyped in 20 members of the CXN family. Marker to marker distances were obtained from the GDB database (http://www.gdb.org/)

Individual:	II:1	III:1	III:2	III:3	III:4	III:5	III:7	III:8	IV:1	IV:2	IV:3	IV:5	IV:6	IV:7	V:1	V:3	V:4	V:5	V:6	V:7
Status:	C	U	U	C	C	C	C	U	A	C	A	A	C	U	A	U	A	U	U	U
Marker																				
DXS1060	?/?	3	3	3/5	3/5	2/3	4/5	1	3	3/5	5	2	5/1	4/1	3	1	5	?	1	2/1
DXS1223	5/5	5	5	3/5	3/5	2/5	2/2	2	3	4/5	5	2	5/2	2/2	4	2	5	2	2	1/2
DXS7108	3/1	3	3	5/1	5/1	3/2	?/?	3	1	1/1	1	3	1/3	4/3	1	3	1	3	4	3/3
DXS9902	2/1	2	2	4/1	4/1	1/4	3/1	3	1	4/1	1	1	3/1	3/3	4	3	1	3	3	3/3
DXS1053	1/2	1	1	3/2	3/2	2/2	3/2	2	2	2/2	2	2	2/2	3/2	2	2	2	?	3	1/2
DXS8036	3/3	3	3	3/3	3/3	3/3	2/3	?	3	2/3	3	3	3/3	3/3	3	3	3	3	3	3/3
DXS1195	1/2	1	1	2/2	2/2	3/2	2/2	2	2	3/2	2	2	2/2	2/2	2	2	2	2	2	1/2
DXS999	2/1	2	2	1/2	1/1	2/1	2/1	2	2	2/2	1	1	2/1	2/2	2	1	1	2	?	?/?
DXS7593	4/5	4	4	3/4	3/5	5/2	1/2	2	4	2/4	4	5	2/2	1/2	4	2	2	2	1	4/2
DXS1061	?/?	3	3	2/3	2/4	4/1	3/2	2	3	3/3	4	4	2/2	3/2	3	2	2	2	3	3/2
DXS1214	2/2	2	2	2/2	2/2	2/4	2/2	1	2	3/2	2	2	2/1	2/1	2	1	2	1	2	2/1
DXS8090	2/2	2	2	5/2	5/2	4/4	2/4	4	2	3/2	2	1	2/4	4/4	2	4	2	4	4	1/4
DXS1068	1/2	1	1	2/1	2/1	2/1	2/1	1	1	2/1	1	2	2/1	1/1	1	1	2	1	1	3/1
DXS8102	1/2	1	1	4/1	4/1	2/4	3/1	3	1	3/1	1	2	3/3	1/3	1	3	3	3	1	2/3
DXS8015	1/1	1	1	4/1	4/1	4/3	3/1	1	1	3/1	1	4	3/4	1/5	1	5	3	5	1	2/5
DXS993	4/2	?	4	1/4	1/4	4/2	?/?	3	4	4/4	4	4	4/3	4/3	4	?	?	3	4	4/3
DXS8080	2/3	2	2	3/2	3/2	3/3	2/2	1	2	3/2	2	3	1/1	1/1	2	1	1	1	1	1/1
DXS8083	1/3	1	1	2/1	2/1	3/1	1/1	2	1	2/1	1	3	1/2	1/2	1	2	2	2	2	1/1
DXS1055	1/2	1	1	4/1	4/1	2/1	1/1	1	4	2/1	1	2	1/1	1/1	1	1	1	1	1	3/1
DXS1039	1/1	1	1	2/1	2/1	1/1	?/?	3	2	2/1	1	1	1/3	1/3	1	3	3	3	3	1/1
DXS991	1/1	1	1	4/1	4/1	2/3	1/1	3	4	5/1	1	2	1/3	1/3	1	3	3	?	3	1/1
DXS1216	3/3	3	3	2/3	2/3	2/3	?/?	3	2	1/3	3	2	3/3	3/3	3	3	3	3	3	2/3
DXS986	2/2	?	2	2/2	2/2	?/?	2/2	3	2	3/2	2	?	2/3	1/3	2	3	3	3	3	2/1
DXS1196	3/4	3	3	2/3	2/3	1/2	1/1	5	2	2/3	3	1	3/5	1/5	3	5	5	5	5	3/1
DXS1217	?/?	2	1	2/1	2/1	2/1	2/1	1	2	1/1	1	2	1/1	1/1	1	1	1	1	1	1/1
DXS8077	2/2	2	2	2/2	2/2	2/2	2/2	2	2	1/2	2	2	2/2	2/2	2	2	2	2	2	2/2
DXS8020	1/4	4	1	1/1	1/1	5/1	4/1	3	1	1/1	1	5	1/3	1/3	1	3	3	3	3	2/1
DXS1106	3/1	?	3	2/3	2/3	3/3	3/3	3	2	3/3	3	?	3/3	3/3	3	3	3	3	3	3/3
DXS1059	?/?	2	1	2/1	2/1	2/1	3/1	3	2	3/1	1	2	1/3	1/3	1	3	3	?	3	2/1
DXS8088	3/3	3	3	4/3	4/3	2/3	1/3	3	4	4/3	3	2	1/3	1/3	3	1	1	1	1	?/?
DXS8055	?/?	?	?	2/3	?/?	1/3	?/?	?	2	1/3	3	1	3/3	3/3	3	?	?	?	3	2/3
DXS8064	3/1	1	3	3/3	3/3	4/3	?/?	3	3	3/3	3	4	3/3	3/3	3	3	3	3	3	2/3
DXS8067	3/4	4	3	2/3	2/3	3/3	4/3	2	2	1/3	3	3	3/2	4/2	3	3	3	2	2	3/4
DXS1001	?/?	2	2	4/2	4/2	2/2	2/2	3	4	1/2	2	2	2/3	2/3	2	2	2	3	3	5/2
DXS8009	2/2	2	2	1/2	1/2	1/2	1/2	2	1	2/2	2	1	2/2	1/2	2	2	2	2	2	2/2
DXS1047	5/4	?	5	1/5	1/5	3/5	?/?	5	1	5/1	5	?	5/5	3/5	1	5	5	?	5	3/5
DXS1062	3/3	3	3	3/3	3/3	2/3	?/?	2	3	2/3	3	2	3/2	1/2	3	2	3	2	2	2/2
DXS984	5/2	2	5	1/5	1/5	4/5	?/?	1	1	1/1	5	4	2/1	6/1	1	1	2	1	1	3/1
DXS1205	1/2	2	1	1/2	1/1	1/1	?/?	2	1	2/1	1	1	2/2	3/2	1	2	2	?	2	2/2
DXS1227	1/1	?	1	1/1	3/1	1/3	1/1	3	3	2/3	1	?	1/3	3/3	3	3	1	3	3	3/3
DXS8106	2/3	3	2	6/3	6/2	4/2	3/3	6	3	1/6	2	4	3/6	3/5	6	6	3	6	6	3/6
DXS8043	4/4	4	4	5/4	5/4	4/4	2/1	3	4	3/5	4	4	1/3	2/3	3	3	1	3	3	1/3
DXS8045	2/2	2	2	2/2	2/2	2/2	?/?	3	2	2/2	2	2	2/3	2/3	2	3	2	3	3	1/3
DXS998	1/2	2	1	3/2	3/1	2/1	3/2	2	2	2/3	1	2	2/2	3/2	2	2	2	2	2	2/2
DXS8069	2/2	2	2	2/2	2/2	1/2	1/2	1	2	1/2	2	1	1/1	1/1	1	1	1	1	1	1/1
DXS1073	2/2	?	2	1/2	1/2	2/3	?/?	2	2	3/1	2	4	3/2	3/2	3	3	3	3	3	3/2

Figure 3.5 Haplotype analysis of the entire X chromosome in the CXN family.

Markers segregating with disease, DXS1053-DXS8036-DXS1195, are indicated (boxed red). NB. Marker DXS1195 was found to be proximal to marker DXS8036 in the current genome assembly, compared to figure 3.1. Positions and distances between markers can be seen in table 3.1. A = affected male, C = carrier female, U = unaffected individual. Where known, both alleles are shown (separated by a virgule [/]) for female members of the CXN family. For individuals position in the pedigree please refer to figure 3.1

The sequences of each of the novel microsatellite markers are shown in figure 3.6. Primers were designed (section 2.2.2) with appropriate fluorescent tags (table 3.2) and microsatellites were amplified using standard conditions (section 2.2.1) with specific annealing temperatures. The products were then loaded and analysed on an ABI 3100 genotyper (section 2.2.7). Genotypes were determined in a double blind-manner. Novel microsatellite markers S6, S7, S8, and S9 (figure 3.6 and table 3.3) proved not to be polymorphic upon genotyping in the CXN family. Of the four novel microsatellites that proved not to be polymorphic, S7 consisted of an imperfect repeat (figure 3.6) whilst the repeat sequences of S8 and S9 (figure 3.6) consisted of only 12 repeats each. S5 (figure 3.6) consisted of 15 tri-nucleotide repeat units and was polymorphic. S7 was identical to S1 in the type and number of repeats (figure 3.6), but only S1 was polymorphic in the CXN family. In summary, five of the nine novel microsatellites (S1-S5) were polymorphic in the CXN family.

Microsatellite	Primer	Sequence (5'-3')	Tm°C	Dye label
DXS1053	Forward	TTAAGGAAGTATGAGGCTCCA	60	6-FAM
	Reverse	TTGGTGCAAAAGTAATCACG	56	6-FAM
S1	Forward	TGTGATACCTGTTATTCACAAAG	62	6-FAM
	Reverse	TATCTGACATTTTAAAGTGCAAG	60	6-FAM
AFMB343YD5	Forward	GTGTCCATTTGTTGCCTACC	60	5-HEX
	Reverse	TGATTTAGATTCCCCAAGTTTTAC	64	5-HEX
S2	Forward	TACATACAGCCCTTCAAAAACTG	64	5-HEX
	Reverse	AGCCAAGACAGGGGGATCAC	64	5-HEX
DXS1195	Forward	TGCAAATGGACCCAAG	48	6-FAM
	Reverse	ATATCACACAGGCACAAATG	56	6-FAM
S3	Forward	TCAGTTCTTGAGTTTCAATTTG	58	6-FAM
	Reverse	TAACCCAGATATTAGAACTAGC	60	6-FAM
S4	Forward	TACTTGAGTGCATCTCCAGG	60	5-HEX
	Reverse	AAAGAGGTACAATTTCAGCCC	60	5-HEX
S5	Forward	ATGTTCATAGCACTTACTATGG	62	5-HEX
	Reverse	AGTGACTTTTAAGGGTATATGG	62	5-HEX
S6	Forward	AAGAGTGTATCTTACTTGGTGC	62	5-HEX
	Reverse	AGAGTTCTGGGAGAAGTTATAG	62	5-HEX
S7	Forward	AGATAATTAGATGAGCTCACCAG	64	6-FAM
	Reverse	TTGTAACCTAGATATGAGTGCTG	64	6-FAM
S8	Forward	ATTTTCTGGCACTAGAATATGC	60	6-FAM
	Reverse	ATCCTGCTTTCTAATACCATTC	60	6-FAM
S9	Forward	AAGTCTATTTCAGTCCAGACAAC	64	5-HEX
	Reverse	AGCGATAACATCAGGAGAAATAC	64	5-HEX

Table 3.2 Additional and novel microsatellite markers identified within the CXN disease interval.

- Novel microsatellite marker S1 (24 di-nucleotide (GT) repeat):

gagtccgtgc tgaaaaggag gatggagaaa cttaggtcaa tacagaccat
TGTGATACCT GTTATTCACA AAGGGACTCT TTCTCATCTT ATGCTAGGAG **S1F**
GGGA**GTGTGT GTGTGTGTGT GTGTGTGTGT GTGTGTGTGT GTGTGTGTGT**
GTTTATTGAG GAGAAATGTG TTGAGAATGA AGAAAGTAAA ACATACATAA
AATATGTAAA AGCATATAAC AATTAGCACA CCGTCTTATA GGTAAAAAGC
TACAATTTTA CAGAAAGGCA ATCTTGCACT TTAAAATGTC AGATAcattt **S1R**
caacaggaca ttgtaaaata gttgtaaaat agtttatctc caggttctca

- Novel microsatellite marker S2 (28 di-nucleotide (GT) repeat):

aggcatctgt gtgttacttt atcttgtttt caactcaaac tttatagttt
TACATACAGC CCTTCAAAAA CTGCTATTTT TTTTTTAAAG ATTTTATTTT **S2F**
TCTTACTGCC TTTTGTTTTT TTTAAAAAAA AATTATGGAA AATTTCAAGC
TTAACAAAAA GTAGAGAGAT AATATAGCAC CTGGCTTCAG CAGATGAAAC
CGTAATTCC**G TGTGTGTGTG TGTGTGTGTG TGTGTGTGTG TGTGTGTGTG**
TGTGTGTGTG TGTGTATAGA GACAGTTTCT TGCTAAATTG CCCAGGCTGG
TCTTGAACTC CTGGGCTCAG **GTGATCCCCC TGTCTTGGCT** ccccaaagtg **S2R**
ctgggattgt agatgtgagc cactgtacat ggctgataat ttcatctctg

- Novel microsatellite marker S3 (24 di-nucleotide (TG) repeat):

attctggtta attttgtcca gtgaatcttt tacttcagat attgtatttt
TCAGTTCTTG AGTTTCAATT TGGCTCCTTT TTATCCTTTG TTTCTCCACT **S3F**
AAGATGGTCT TTTTAATCAG TACGAT**TGTG TGTGTGTGTG TGTGTGTGTG**
TGTGTGTGTG TGTGTGTGTG TGTGTCTGTG TAATGTCATC GAAGACCAGT
TGAATAGCCA CTTTAAAATT TTTGTTT**GCT AGTTCTAATA TCTGGGTTAT** **S3R**
acttgggtct gtctcagttg attttctttt ttcttgaaaa tattattcat

- Novel microsatellite marker S4 (21 di-nucleotide (GT) repeat):

agcctggtcc ctttggaagc agtcacgtgg ccgctgactt tgcagcccat
TACTTGAGTG CATCTCCAGG CCAGGAACAG AGCAGGGCTG GGGCAGAGAG **S4F**
CTGTGCTCAC GGCCAGTGCA TTGTCTCCCA GATTCGTGGC TCTGGGCTTG
CCTTCCCGGC TGAAAGGCAA AGCAGCCTGG CCCTGTTCTA GTTATTTCTT
TGTGTATATA TGTTTGTTTA TTTATTTAAG CAAGAGCTCT CTCTGCAGAG
TGGG**GTGTGT GTGTGTGTGT GTGTGTGTGT GTGTGTGTGT GTGTGT**TTGT
GTGGGGTGCT TCAGAGA**GGG CTGAAATTGT ACCTCTTT**TC TTTCAGCCCT **S4R**
ctgcatttcc ccagagagaa ggtttttttc tcgtcttcat ttcctttgaa

- Novel microsatellite marker S5 (15 tri-nucleotide (TTA) repeat):

ctataatctt accaagactg aatgaaagta tttggtgctt atacttttta
ATGTTCATAG CACTTACTAT GGCCAGGCAC AATGCCCAAG TGTTTTGCT**T** **S5F**
TATTATTATT ATTATTATTA TTATTATTAT TATTATTATT ATTATTTTCA
CAACAACCCA GCGAAACAGG TACTATTATT ATCCCCTTTT TGTAAATGAG
GAAACAGATT TCCATACTTG ACATTAACCT GCCCAAGGTC ACCCAGTTTG
GAAGTGGCAG GGCCAGGATT TGAAGCCGAA TATGTTCATT GTAAAGC**CCA** **S5R**
TATACCCTTA AAAgtcactg cacgtccacg tatggtaatg tggtcacaat

- Novel microsatellite marker S6 (22 di-nucleotide (GT) repeat):

```
ATAATTTATT CTGATTTTTC TGGCTCACAT ATTGTGTCCC CAGTAGAAAC
AAGAGTGTAT CTTACTTGGT GCCATCGAAG TCTGTCCAAG TTTGAAGGGG S6F
TGTGTGTGTG TGTGTGTGTG TGTGTGTGTG TGTGTGTGTG TGTAGGGGAG
ACAACCTCAT GTGCTCATTC TCTATAACTT CTCCCAGAAC TCTtgtttgt S6R
tgaattgatt gtaccttgag agatgctttc ccctttcaca cacctacctc
```

- Novel microsatellite marker S7 (25 di-nucleotide (GT) repeat):

```
gtagaaatac aaatttgaaa atcatttgca tataaactgt atttaaagcc
AGATAATTAG ATGAGCTCAC CAGGGGATTA ATACAGGTTA GGGTTTCTCA S8F
GCCTTGGCAC TACTGACGTT TGGAGCCAGA ATGGGTGTGC GTGTGTGTGT
GTGTGTGTGT GTGTGTGTGC GCGTGTGTGT GTGTATGTGT ATTCTGTGCA
TTGTAGGATG TGAAGCAGCA CTCATATCTA GGTTACAAag catcatttca S8R
ctactctctt ctcagcatcc aaaaccaact tattcccttg tctttgccct
```

- Novel microsatellite marker S8 (12 tetra-nucleotide (TTTC) repeat):

```
cctcctcctc ctccttctct ctctctcttt ttttttaaac ctacttcttt
ATTTTCTGGC ACTAGAATAT GCTCCAGGCT CATCTTGTAC ATTTTTATGC S11F
CCCAGTCCTA GTTTCAGCCA TTTCACCAAA GAGTCCTGGT TTCTTTCTTT
CTTTCTTTCT TTCTTTCTTT CTTTCTTTCT TTCTTTCTTT TGAGGCAGGG
CGTCACTCTC ACCCAGGCTG GAGTGCAGTT GCATGATCGC GGCTCACCAC
AGCCTTGACC TCCTAGGCTC AGGTGATCCT CCCACCTCAG CCTCCAGAGT
AACTGGGGCT ACAGGTGCGG GCCACCACAC CTGGCTAATT TTTGTATTTT
TTTTTTTTTT TTGTAGAGAA GGGGTTTTAC CATGTTGCCC AGGCTGGTCT
CGAACTCTTG GGCTCAAGTG ATCTGCCCGG CCTGTCTCCC AAAGTGCTAG
GATTACAGAC ATGAGCCACT GTGCCCGACT CCTGGTTTAT TTTTATCACA
GAATGGTATT AGAAAGCAGG ATctaggtgc taggtgggtg ctaatgaatc S11R
tttcattttg gttatttact tt
```

- Novel microsatellite marker S9 (12 tri-nucleotide (TTA) repeat):

```
ttttttttt ttgtattaat gaaacgaaag tttatttctt gctcatacaa
AAGTCTATTT CAGTCCAGAC AACTCCACAG GGAAATTGTC CTTCATGTAA S12F
TAGATCAGTA TTCCAGGTTA CTTAGTTATT ATTATTATTA TTATTATTAT
TATTATTATT ATGCTTTAAG TTCTGTGTTA CATGTGCAGA ACATGCAGTT
TTGTTACATA GGTATACAGG TGCCATGGTG GTTTGCTGCA CCCATCAACC
CATCACCTAC ATTAGGTATT TCTCCTGATG TTATCGCTCC CCTAGCCCCC S12R
caccccacag aggccccagt gtgtgatgtt cccctccctg tgaccatgtg
```

Figure 3.6 Sequences of novel microsatellite markers S1, S2, S3, S4, and S5.

The nucleotide repeat sequences are highlighted in green with primer sequences shown in red. The imperfect repeats in microsatellite marker S7 are underlined in green.

Figure 3.7 shows examples of the ABI 3100 traces obtained for microsatellite marker AFMB343YD5 in a carrier female, her affected son, and an unaffected individual in the CXN family.

Novel microsatellite marker	Type of repeat	Repeat sequence	Number of repeats
S1	Di-nucleotide	GT	24
S2	Di-nucleotide	GT	28
S3	Di-nucleotide	TG	24
S4	Di-nucleotide	GT	21
S5	Tri-nucleotide	TTA	15
S6	Di-nucleotide	GT	22
S7*	Di-nucleotide	GT	25
S8	Tetra-nucleotide	TTTC	12
S9	Tri-nucleotide	TTA	12

Table 3.3 Summary of all novel microsatellite markers genotyped in the CXN family

* denotes an imperfect repeat sequence

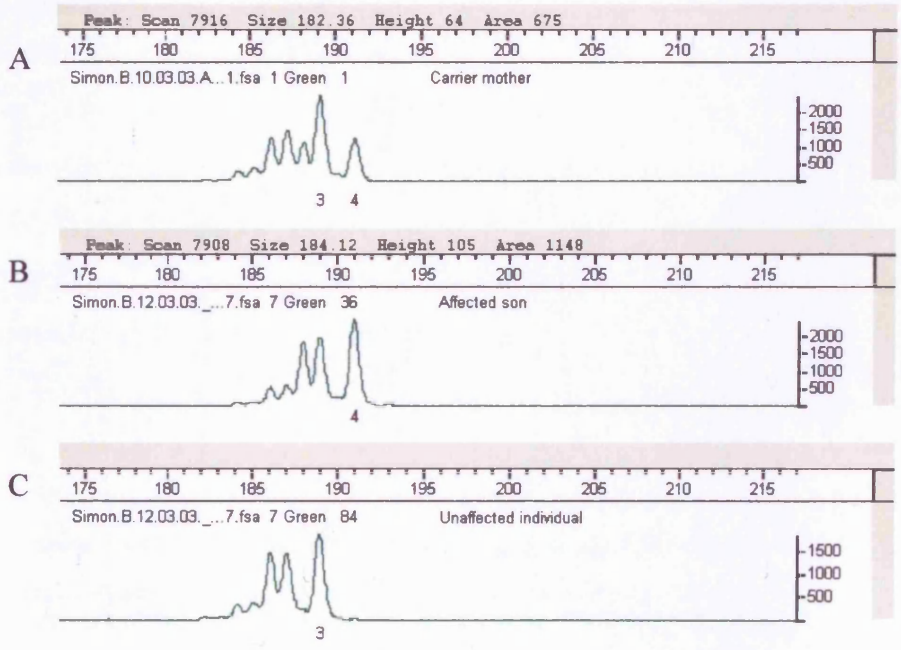


Figure 3.7 ABI 3100 genotyping results for microsatellite marker AFMB343YD5.

Genotyping results obtained for a carrier female (A) and her affected son (B) in the CXN family compared to an unaffected individual (C) within the same family. The unaffected individual is homozygous for allele 3, whilst the affected son is hemizygous for allele 4 and the carrier female heterozygous for alleles 3 and 4. Allele 4 segregates with disease in the CXN family.

Figure 3.8 depicts the pedigree with new haplotypes, indicating relative genetic positions of all microsatellites typed in relation to previously identified recombinant markers (DXS9902 and DXS999) and linked markers (DXS1053 and DXS1195). The reconstructed haplotypes confirm disease segregation with markers in Xp22.13-22.2 (DXS1053-0.8Mb-S1-1.23Mb-AFMB343YD5-0.1Mb-S2-1.07Mb-DXS1195-0.84Mb-S3, figure 3.8). The proximal cross over was originally identified between markers DXS8036 and DXS999 (figure 3.1). Marker DXS1195, which is linked with disease has now been localised proximal to DXS8036 after confirming relative positions of microsatellite markers in the latest genome build (UCSC, http://genome.ucsc.edu/). Novel microsatellite markers S4 (within intron 2 of *PPEF1*) and S5 (within intron 2 of *CDKL5*) were both found to be crossed over in individual III.3 and subsequent generations (IV:1, IV:2, and V:1) thereby refining the proximal recombination to between S3 and S5 (figure 3.8). A recombination event defines the distal cross over between DXS9902 and DXS1053 (seen in individual V:1). The disease interval has therefore been refined to an approximately 3.2 Mb interval between markers DXS9902 and S5 (figure 3.8).

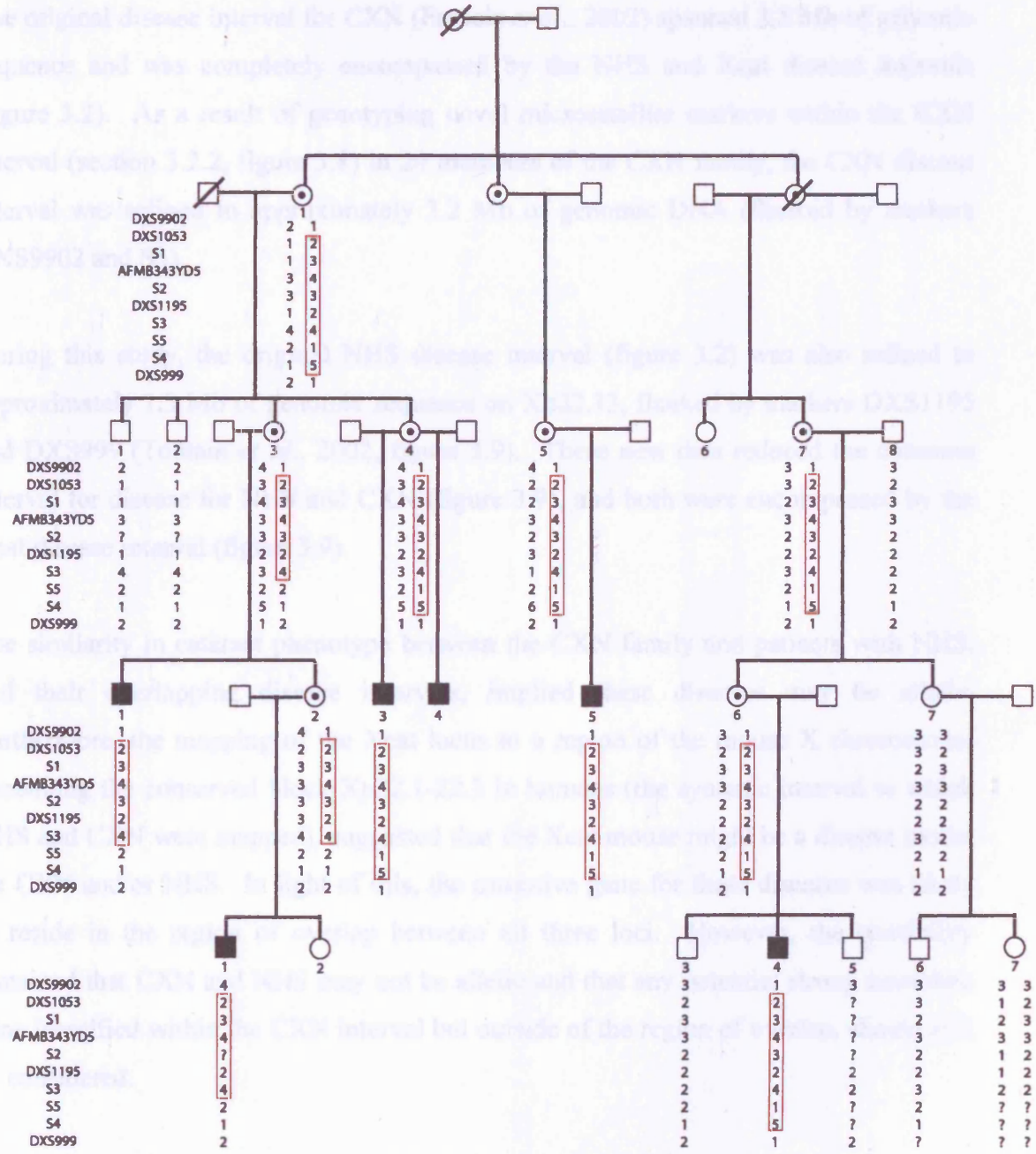


Figure 3.8 CXN pedigree showing haplotypes on Xp22.13 with new microsatellites.

Boxed haplotypes of affected males and carrier females show recombination with markers S4, S5, DXS999, and DXS9902. Disease segregates with markers DXS1053-S1-AFMB343YD5-S2-DXS1195-S3

3.2.3 Candidate genes for CXN and NHS

The original disease interval for CXN (Francis *et al.*, 2002) spanned 3.5 Mb of genomic sequence and was completely encompassed by the NHS and Xcat disease intervals (figure 3.2). As a result of genotyping novel microsatellite markers within the CXN interval (section 3.2.2, figure 3.8) in 20 members of the CXN family, the CXN disease interval was refined to approximately 3.2 Mb of genomic DNA (flanked by markers DXS9902 and S5).

During this study, the original NHS disease interval (figure 3.2) was also refined to approximately 1.3 Mb of genomic sequence on Xp22.13, flanked by markers DXS1195 and DXS999 (Toutain *et al.*, 2002, figure 3.9). These new data reduced the common interval for disease for NHS and CXN (figure 3.9), and both were encompassed by the Xcat disease interval (figure 3.9).

The similarity in cataract phenotype between the CXN family and patients with NHS, and their overlapping disease intervals, implied these diseases may be allelic. Furthermore, the mapping of the Xcat locus to a region of the mouse X chromosome containing the conserved block Xp22.1-22.3 in humans (the syntenic interval to which NHS and CXN were mapped), suggested that the Xcat mouse might be a disease model for CXN and/or NHS. In light of this, the causative gene for these diseases was likely to reside in the region of overlap between all three loci. However, the possibility remained that CXN and NHS may not be allelic and that any potential strong candidate gene identified within the CXN interval but outside of the region of overlap, should still be considered.

Twenty-two known genes (i.e., named and/or characterised at some level, as opposed to predicted, unconfirmed genes) were found to reside in the CXN interval at the outset of this project (table 3.4) based on the current genome database (genome version January 2003 freeze). Three of the 22 known genes resided within the region of overlap between the CXN and NHS intervals: *sex comb on midleg-like-1* and *-2* (*SCML1*, *SCML2*), and *Retinoic acid induced 2* (*RAI2*, figure 3.9).

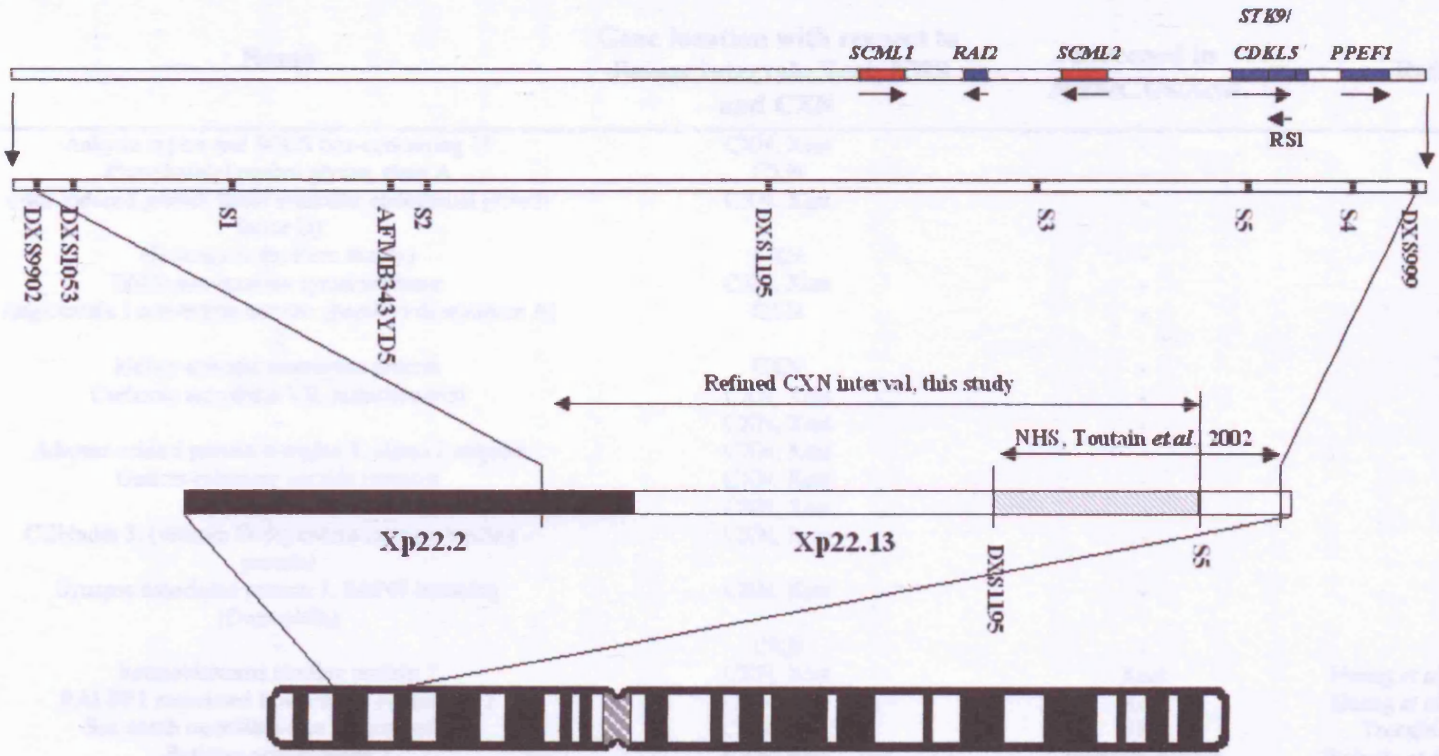


Figure 3.9 Revised schematic diagram of the refined genomic intervals for CXN and NHS.

Genomic intervals shown above are based on recent refinements for NHS (Toutain *et al.*, 2002) and CXN (this study). Genes boxed blue have been excluded for CXN (this study), and NHS (Walpole *et al.*, 1999; Toutain *et al.*, 2002) whilst those boxed red were excluded for NHS (Toutian *et al.*, 2002). The grey shaded area within the genomic region Xp22.13 denotes the region of overlap between the CXN and NHS intervals.

Gene	Name	Gene location with respect to disease intervals Xcat, NHS and CXN	Screened in *NHS/CXN/Xcat*	Reference
ASB11	Ankyrin repeat and SOCS box-containing 11	CXN, Xcat	-	-
PIGA	Phosphatidylinositol glycan, class A	CXN	-	-
FIGF	c-fos induced growth factor (vascular endothelial growth factor D)	CXN, Xcat	-	-
PIR	(Belongs to the Pirin family)	CXN	-	-
BMX	BMX non-receptor tyrosine kinase	CXN, Xcat	-	-
ACE2	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 2	CXN	-	-
NX17	kidney-specific membrane protein	CXN	-	-
CA5B	Carbonic anhydrase VB, mitochondrial	CXN, Xcat	-	-
U2AF1-RS2	-	CXN, Xcat	-	-
AP1S2	Adaptor-related protein complex 1, sigma 2 subunit	CXN, Xcat	-	-
GRPR	Gastrin-releasing peptide receptor	CXN, Xcat	-	-
CTPS2	CTP synthase II	CXN, Xcat	-	-
CALB3	Calbindin 3, (vitamin D-dependent calcium binding - protein)	CXN, Xcat	-	-
SYAP1	Synapse associated protein 1, SAP47 homolog (Drosophila)	CXN, Xcat	-	-
CXORF15	-	CXN	-	
RBBP7	Retinoblastoma binding protein 7	CXN, Xcat	Xcat	Huang *et al.*, 2003 ARVO
REPS2	RALBP1 associated Eps domain containing 2	CXN, Xcat	Xcat	Huang *et al.*, 2003 ARVO
SCML1	Sex comb on midleg-like 1 (Drosophila)	CXN, NHS	NHS	Toutain *et al.*, 2002
RAI2	Retinoic acid induced 2	CXN, NHS	NHS/CXN/Xcat	Walpole *et al.*, 1999; Huang *et al.*, 2003 ARVO;
SCML2	Sex comb on midleg-like 1 (Drosophila)	CXN, NHS, Xcat	NHS	Toutain *et al.*, (2002)
CDKL5	Cyclin-dependent kinase-like 5	CXN, NHS	NHS	Toutain *et al.*, (2002)
RS1	Retinoschisis (X-linked, juvenile) 1	CXN, NHS, Xcat	NHS	Toutain *et al.*, (2002)
PPEF1	Protein phosphatase, EF hand calcium-binding domain 1	CXN, NHS	NHS	Toutain *et al.*, (2002)

Table 3.4 Characterised genes within the CXN, NHS and Xcat loci.

All three genes (*SCML1*, *SCML2*, and *RAI2*) were reported to have been excluded as disease causative for NHS (Toutain *et al.*, 2002). The *RAI2* gene had also been screened and excluded for CXN (Francis, unpublished data at the time). An additional two genes were reported to be excluded for NHS: *protein phosphatase with E-F hand motif* (*PPEF1*), *retinoschisis (X-linked, juvenile) I* (*RS1*), *serine-threonine kinase 9* (*STK9/CDKL5*; Toutain *et al.*, 2002). All three of these genes (*PPEF1*, *RS1* and *STK9/CDKL5*) have been excluded for CXN through the refinement of the CXN interval (section 3.2.2, figure 3.9) Due to all three characterised genes (*SCML1*, *SCML2*, and *RAI2*) within the region of overlap being excluded for either CXN, NHS, or both and since no gene predictions were found to reside in the region of overlap at the time, candidate genes within the entire CXN interval were sought for screening in the CXN family.

3.2.3.1 Identification, characterisation, and mutation screening of candidate genes

Identification of candidate genes within the CXN and NHS disease intervals were sought through a bioinformatic approach (section 2.7). A powerful strategy for identifying candidate genes in this study was through comparative genomics of the syntenic region on the mouse X chromosome containing the Xcat locus (figure 3.10). Gene prediction programmes (Twinscan and Syntenic Gene Prediction, SGP) based on human-mouse homology were used to identify novel candidate genes, and were analysed using NIX and PIX analysis programmes to determine any homologies to known genes/proteins, respectively.

Four candidate genes; one fully characterised gene (*RBBP7*, table 3.4) and three novel genes-*TL1*, *CX43-L* and *NCG1*) were identified and screened for mutations. *RBBP7* and *TL1* resided within the CXN and Xcat intervals, whereas *CX43-L* and *NCG1* mapped within all three disease loci.

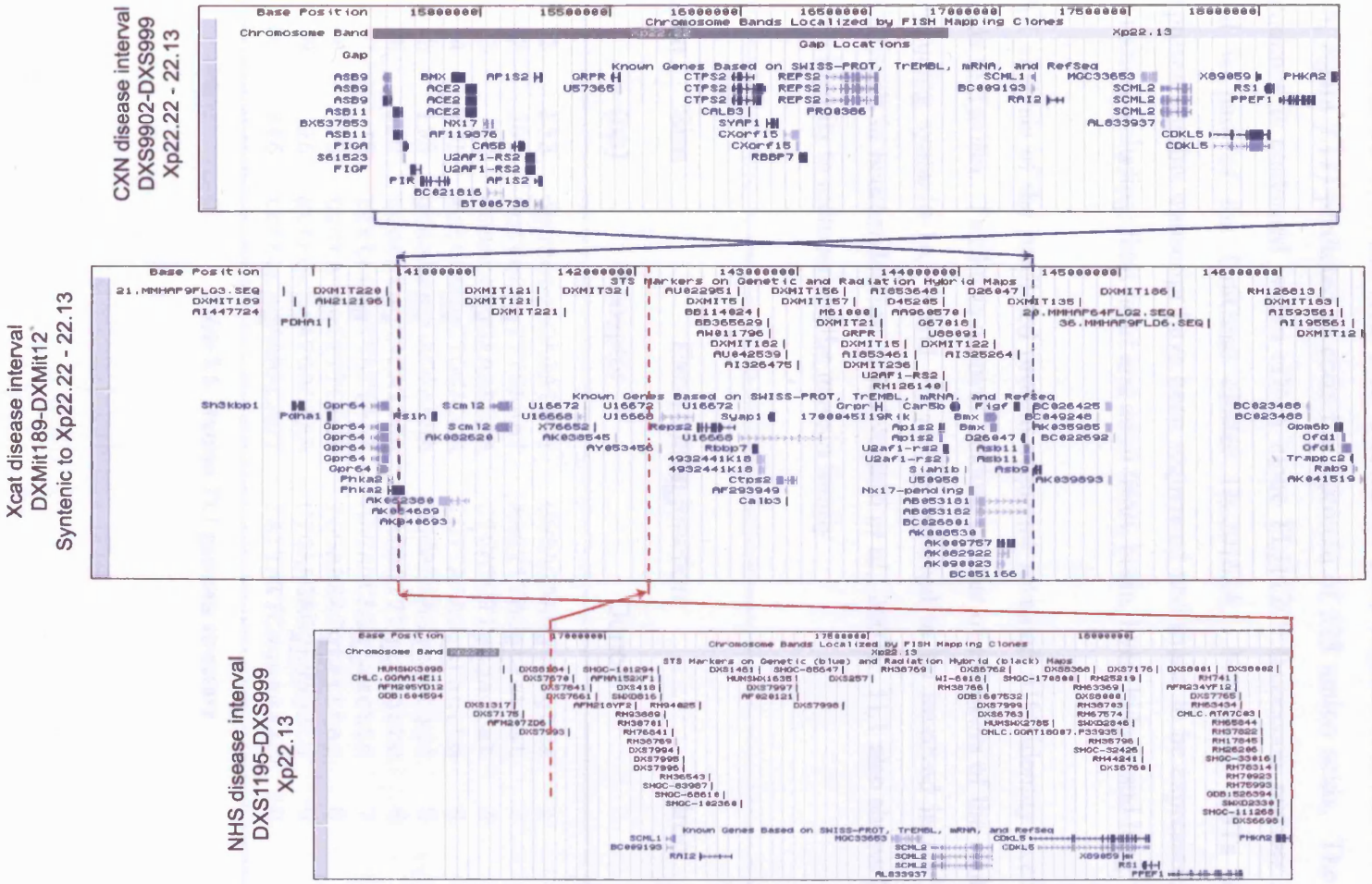


Figure 3.10 Alignment of the human CXN, NHS, and mouse Xcat disease intervals.

Known and uncharacterised genes within the intervals are shown. Taken and adapted from the UCSC genome browser 2003 Freeze (http://genome.ucsc.edu/cgi-bin/hgGateway)

3.2.3.1.1 Taxilin-like 1 (TL1)

A novel candidate gene (CXorf15, hereafter named *Taxilin-like 1*, *TL1*) predicted by gene prediction programmes in both human and mouse genomic sequence was found to reside in the CXN and Xcat loci. The human gene consists of at least 10 exons (table 3.5, figure 3.11) predicted to code for a protein of 528 amino acids. The full coding sequence is contained within cDNA clone FLJ11209 (accession number AK002071) and is part of the UniGene cluster Hs.201624. Numerous ESTs and cDNAs representing this transcript have been sequenced and found to be expressed in a variety of tissues including: fetal heart and adult heart, brain, liver, kidney and lung.

PIX analysis of the predicted protein sequence revealed 61% identity over 327 amino acids to Taxilin. Taxilin is a novel binding partner of members of the syntaxin family (including syntaxin-1a, -3 and -4) and is believed to be involved in Ca^{2+}-dependent exocytosis in neuroendocrine cells (Nogami *et al.*, 2003). TL1 also shows between 20-30 % identity to members of the myosin family

Exon	Size	Exon-intron junction		Intron	Size
	(bp)	Acceptor	Donor		(bp)
1	153	ggccctgcGAGATTCT	GGCAGAAG**gt**cagtca	1	31984
2	304	tcttct**ag**TTTGAAAT	AACTTTAG**gt**aataat	2	1284
3	92	accatc**ag**GAAAAGAA	CTGATCTT**gt**gagtat	3	7836
4	171	aattttagCTGGAGGA	CGTTAAAG**gt**tagttg	4	1311
5	195	gtaata**ag**GAGGAAAA	GGGAAGAG**gt**aatggc	5	2852
6	120	tcccat**ag**CACATTGA	GAGAGTTT**gt**aagttc	6	1540
7	75	ttttct**ag**TTATTAAA	AACAGCAG**gt**aacttc	7	3235
8	93	tcctat**ag**CTTTCTCT	TGGAAAAG**gt**atttac	8	2135
9	96	attcac**ag**ATGACAAA	CTGAAGAG**gt**gagatg	9	1511
10	746	tcttac**ag**AAAACAGT	TATTATTCagaaaaac	10	

Table 3.5 Human *TL1* genomic structure

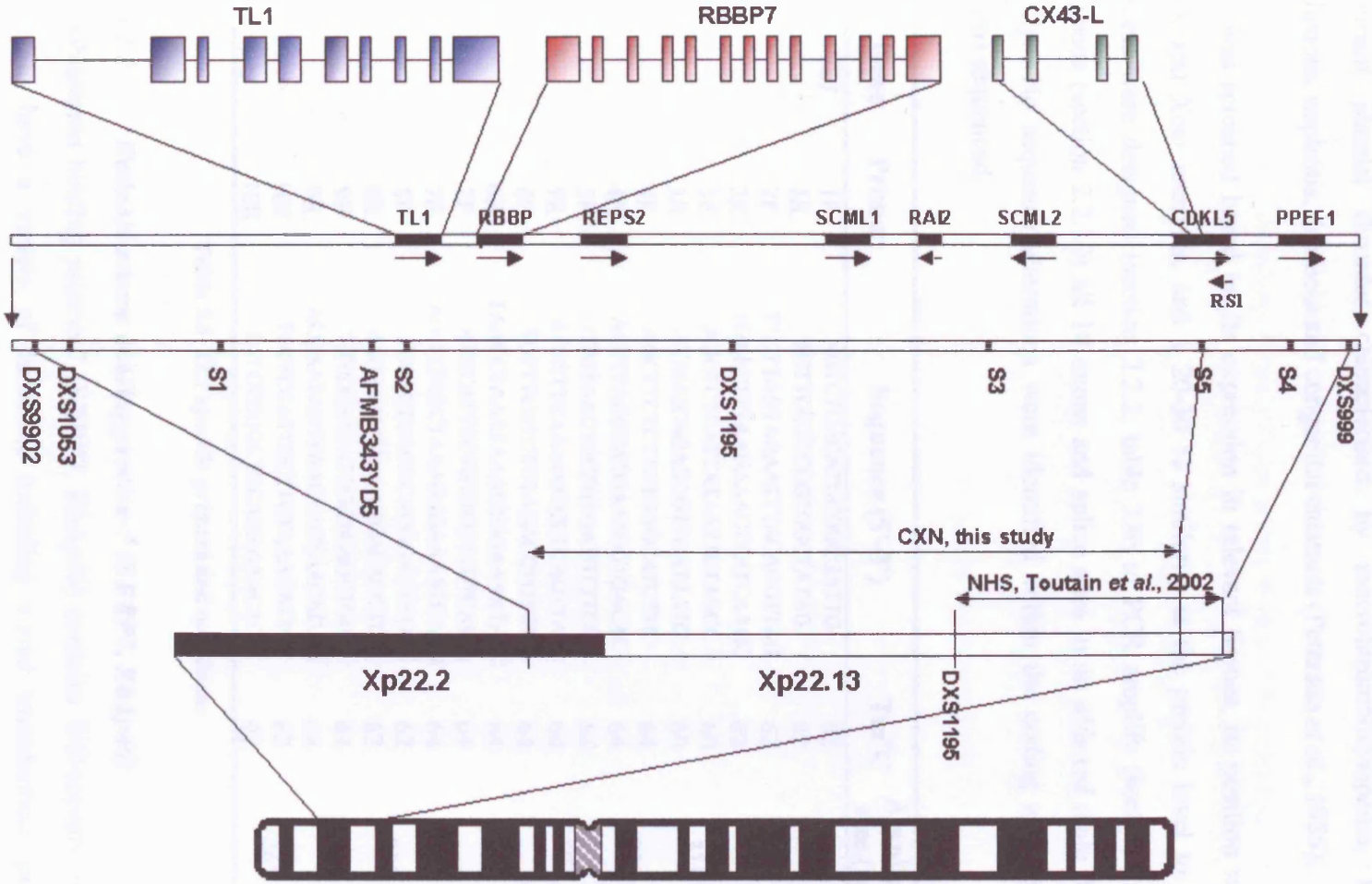


Figure 3.11 Schematic diagram of the genomic interval for CXN and NHS on Xp22.13.

Relative positions of excluded genes *PPEF1, CDKL5, TL1, RBBP7* and *RAI2* are shown with the current microsatellite map.

Mutations within *nonmuscle myosin heavy chain 9* (*MYH9*) gene have been identified in Fechtner syndrome (FTNS, MIM #153640; May-Hegglin/Fechtner Syndrome Consortium 2000). Peterson and colleagues first described FTNS as an autosomal-dominant platelet disorder characterised by macrothrombocytopenia, leukocyte inclusions, nephritis, deafness and congenital cataracts (Peterson *et al.*, 1985).

TL1 was screened based on its expression in relevant tissues, its position within the CXN and Xcat intervals, and a 20-30 % similarity at the protein level to myosins. Primers were designed (section 2.2.2, table 3.6) to PCR amplify (section 2.2.1) and sequence (section 2.2.10) all 10 exons and splice sites in an affected male from each family. No sequence alterations were identified within the coding and non-coding regions sequenced.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
TL1	1F	AGTCTCACATGCGGCGATTG	62	320
	1R	TGTTCGTCCCGTGGCTATAG	62	
	2F	TTGTTAGTAGAATTTACAGGTTAG	62	245
	2R	TGTACTTTAAGAAACTTATCAAGC	62	
	3F	AGCTCTCATTATAATTCTAGCC	60	318
	3R	ATAGGCACAGGGTTGATTATG	60	
	4F	AGCTTCTCTTGTGAGCATCTTC	64	330
	4R	ACTTTAGGTATGAATGGTGACAC	64	
	5F	CTGTAACTGCTGTGATGTTTCC	64	265
	5R	ATGTTTCAAATGGCCTCAGTTAC	64	
	6F	TGTTTCATCTGTAGACCTGTGG	64	377
	6R	TAACCTAAATAAAGTGGAAAATACC	64	
	7F	ATGCATTTCTGTGGGTTGTCAG	64	398
	7R	AGGATCACTAAAGAGAAAATGAAG	64	
	8F	TGTACTGGGACAGGGAGTAG	62	260
	8R	AGTAGTGTTACAGCACACGTC	62	
	9F	TGTCCAATATGGCTGAGCTGG	64	496
	9R	ACCAATAGTTAACTACTAACAGAG	64	
	10F	TGTTCCATTGCTTGTAAACATTG	62	463
	10R	TATCTGGGCTGCAGAAAACAC	62	

Table 3.6 *TL1* specific primers and conditions

3.2.3.1.2 Retinoblastoma binding protein-7 (RBBP7, RbAp46)

Retinoblastoma binding protein-7 (RBBP7, RbAp46) contains WD-repeats, which are known to have a variety of functions including signal transduction, pre-mRNA processing and cytoskeleton assembly (Qian and Lee, 1995). RBBP7 is a core component of at least two major co-repressor complexes, and is thought to have a role modulating transcriptional activity and chromatin remodelling (Vermaak *et al.*, 1999;

Parthun *et al.*, 1996; Verreault *et al.*, 1998; Ahringer, 2000). This 425 amino acid protein is highly conserved between human and mouse and is ubiquitously expressed. *RBBP7* is expressed from embryonic day 9.5 through to adulthood and is detected in the developing brain, heart and eye amongst other tissues (Yang *et al.*, 2002). The genomic structure of the gene (figure 3.11) is also conserved between human and mouse containing at least 12 coding exons. Since genes involved in chromatin remodelling have been implicated in developmental syndromes (e.g. the *ribosomal protein s6 kinase, RSK2,* in Coffin-Lowry syndrome), and *RBBP7* maps within the disease intervals for CXN and Xcat, it was considered a candidate for disease. Primers were designed (section 2.2.2, table 3.7) to PCR amplify (section 2.2.1) and sequence (section 2.2.10) all 12 exons and splice sites in an affected male from each family. No nucleotide changes were identified in all coding and non-coding regions sequenced.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
RBBP7	1F	TCTTGGCTAACGAGAGGAGTC	64	488
	1R	ACGCCAATTCGCGCCTTTCG	64	
	2F	TTCTGACTTTGATCTATCTCAGC	64	263
	2R	AACTAGGAAATTCAGAACGACAG	64	
	3F	ACAGTTGATGCTAAGTATATATTTG	64	388
	3R	TAAAGTCCTAGAAGTTAAATTACTG	64	
	4F	TATTTTCCAGTGCAAATGGGC	64	520
	4R	TTGAGGATGGTTATGGGACC	55	
	5F	TTCTCACCAGTTCTGATAAACAG	55	247
	5R	TATAGTACTCCTTCAGCACAGC	64	
	6F	CAGTCTCTTATAGCCCTTGTTC	64	265
	6R	TGATGGTCATTCAGGTCACCC	64	
	7F	TGCAGTGACAGTGATGTGAGG	64	266
	7R	ATCTGATTAAGGCACAAGAGAAC	64	
	8F	TTGAAAATTCTTAGGTAAAGCATAG	64	182
	8R	TCCCTGGATTTCATTTCCTGAC	64	
	9F	TTGAAAGTCGGGGACAAGAG	60	360
	9R	ACAGAACAAAGGCTCTTTAGG	60	
	10F	TGTTCGTGCTACAAGAATTTCAC	64	171
	10R	AAAGAGATCAATGACAATCCAATC	64	
	11F	AATAGATTACTGAGCCTGTACC	62	
	11R	ATGCAGAGATTTCTCCTGACC	62	488
	12F	ACATGACCATGTGAAAACCATAG	64	
	12R	AATCAAGCATTCATGTAGCATTAC	64	184

Table 3.7 *RBBP7* specific primers and conditions

3.2.3.1.3 Connexin 43-like (CX43-L)

Connexin proteins are major components of gap junctions, which mediate intercellular transportation of small biomolecules (<1 kDa) such as second messengers, ions, and metabolites. Connexin 43 is thought to play an important role in the synchronized contraction of the heart and in embryonic development (Britz-Cunningham *et al.*, 1995).

Paznekas and colleagues recently reported the finding that mutations within the connexin 43 (*GJA1*) gene, cause the pleiotropic phenotype of Oculodentodigital Dysplasia (ODDD [MIM 164200]; Paznekas *et al.*, 2003). ODDD is an autosomal dominant disorder with intra- and interfamilial variability. The ODDD phenotype has many features in common with either Nance-Horan Syndrome or X-linked congenital cataract. Ophthalmic findings include cataract, microphthalmia, and microcornea. Most cases of ODDD have abnormal primary and permanent dentition, craniofacial anomalies and cleft palate. In addition, some families present mild mental retardation and cardiac abnormalities are observed on occasion (Paznekas *et al.*, 2003).

Located in the disease intervals for CXN and NHS is a predicted connexin 43 pseudogene. The pseudogene has 82.1 % sequence identity at the nucleotide level to connexin 43 and resides within intron 3 of the *Cyclin-dependent kinase-like 5* (*CDKL5*) gene, but does not contain a single continuous open reading frame. The Xcat locus does not contain a connexin 43 pseudogene, but does contain a connexin 33 gene (*Gja6*).

Bioinformatic analysis of this genomic region containing the CX43 pseudogene led to the identification of a predicted gene (here named *CX43-Like*, *CX43-L*, figure 3.11) based on two gene prediction programmes (Genscan and the SGP), which shows an 82 % identity to *CX43*. No EST or cDNA hits were found to align with this newly predicted four exon gene. In addition to these four exons forming a novel gene prediction, it was possible that these exons (except exon 2) may prove to be cryptic exons of the *CDKL5* gene. Primers were designed (section 2.2.2, table 3.8) to PCR amplify (section 2.2.1) and sequence (section 2.2.10) the predicted exons in an affected male from each family due to its location within the CXN and NHS intervals and its similarity to connexin 43. No nucleotide changes were identified in the coding and non-coding regions sequenced in either of the affected males. These exons were screened prior to obtaining the genotyping results in which microsatellite marker S5

recombined with disease (section 3.2.2), therefore leading to the exclusion of *CDKL5* as a candidate for disease.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
CX43-L	1F	AGATACCACGCATTTGCTGTC	62	262
	1R	ATTCCAAAGAGATCTGAGAAGC	62	
	2F	AGGGAGTCATTTAATACTTCATG	62	285
	2R	AGGAATAACTAACTGTTCATTGC	62	
	3F	ACCAGTCTTTGCCAATCTCC	60	291
	3R	AAGATGACACCAATGGTACAC	60	
	4F	ATATTGCAAGACTCTGGATCC	60	364
	4R	ATTCCCTGCCTGATGAGATC	60	

Table 3.8 *CX43-L* specific primers and conditions

3.2.4 Discussion

The CXN locus was originally mapped to a 3.5 Mb interval on Xp22.13-22.2 (Francis *et al.*, 2002). Segregation of disease to this genomic region was confirmed in this study by genotyping 46 known microsatellite markers spanning the human X chromosome. An attempt was then made to further refine the genomic interval in order to reduce the number of possible candidate genes for screening in the CXN family. Novel microsatellites markers were identified and genotyped in house. This resulted in the identification of a crossover in individual III.3 and subsequent generations with novel markers S4 and S5. The disease interval has therefore been refined to an approximately 3.2 Mb interval flanked by markers DXS9902 and S5. Microsatellite markers S4 and S5 were identified within intron 2 of *PPEF1*and *CDKL5* respectively. Therefore, the refinement of the CXN interval led to the exclusion of *PPEF1*, *RS1* and *CDKL5* as candidates for disease in the CXN family.

The refined CXN interval overlaps the recently refined NHS disease interval, which spans approximately 1.3 Mb of genomic sequence flanked by markers DXS1195 and DXS999 (Toutain *et al.*, 2002). If CXN and NHS are allelic then refinement of the CXN disease interval implies the disease causing gene now lies within approximately 1 Mb of genomic sequence flanked by markers DXS1195 and S5 (figure 3.11). Further, exclusion of the *PPEF*, *RS1* and *CDKL5* genes in the CXN family is in agreement with the exclusion of these genes as causative for NHS (Toutain *et al.*, 2002).

The possibility that CXN and NHS may not be allelic could not be overlooked, such that potential candidate genes for CXN were also sought outside the region of overlap. Two genes residing outside the interval of overlap, which were present in the Xcat locus, were *TL1* and *RBBP7*. Both genes were thought to be strong candidates for disease in the CXN family and were screened for mutations. No nucleotide alterations were identified in the coding and non-coding regions sequenced for either *TL1* or *RBBP7* resulting in their exclusion as disease causative for CXN. A third candidate gene (*CX43-L*) based on a gene prediction and displaying homology to CX43 was screened in the CXN and NHS families but no sequence variants were identified in both coding and non-coding regions sequenced. The *CX-43-L* gene prediction resided within intron 2 of the *CDKL5* gene, which was excluded as disease causative for CXN through refinement of the CXN interval. Consequently, the refinement also excluded the *CX43-L* gene as a candidate. The total number of genes now excluded for CXN from this study is 6 (*PPEF1*, *RS1*, *CDKL5*, *TL1*, *RBBP7*, and *CX43-L*). In addition, the *RAI2* gene was previously excluded as disease causative for CXN.

To date a total of 6 genes have been reported to have been screened and excluded as disease causative in NHS (*RBBP7*, *CDKL5*, *PPEF1*, *RAI2*, *RS1*, *SCML1* and *SCML2*; this study; Toutain *et al.*, 2002; Walpole *et al.*, 1999). Within the region of overlap between the refined disease loci for CXN and NHS, all known genes (*SCML1*, *RAI2*, and *SCML2*) have been screened and excluded for either CXN or NHS. Only *SCML1*, *SCML2* have not been screened in the CXN family.

The mouse Xcat locus is syntenic with and completely encompasses the refined CXN and NHS loci. X-linked congenital cataract (Xcat) in the mouse displays a similar phenotype to that observed in CXN and NHS. Based on these observations, the Xcat mouse may prove to be a model of disease for either CXN and/or NHS. Recently, seven genes (*Rai2*, *Rbbp7*, *Ctps2*, *Calb3*, *Grpr*, *Reps2*, and *Syap1*) residing in the Xcat locus were reported to have been sequenced in the Xcat mouse and no mutations were identified (Huang *et al.*, 2004). All seven genes reside within the CXN interval, but only *RAI2* resides within the NHS interval and therefore within the region of overlap between CXN, NHS, and Xcat. If the Xcat mouse is a disease model for CXN, then the exclusion of *RBBP7* in both CXN and Xcat supports this possibility, as does the

exclusion of *RAI2* in all three diseases. The *RAI2* gene is currently the only gene screened and excluded for all three diseases (CXN, NHS and Xcat).

All known genes within the refined NHS locus have now been screened and excluded for NHS. This would imply that an uncharacterised gene residing in the NHS locus remains to be identified. The likelihood that CXN and NHS are not allelic, cannot be ignored and future candidate gene screening for CXN will be considerable due to the relatively large interval (currently 3.2 Mb of genomic sequence). The acquisition of new families with X-linked congenital cataract may help to refine the interval further and therefore reduce the size of the CXN locus in which candidate genes reside.

A recent publication by Ng and colleagues (2004) reported that distinct classes of mutations within the *BCOR* gene on Xp11.4 (encoding BCL-6-interacting corepressor) result in the Oculofaciocardiodental (OFCD, MIM 300166) and Lenz microphthalmia syndromes. OFCD is a distinct form of syndromic microphthalmia associated with congenital cataracts, narrow face, broad nasal tip with separated cartilage, cleft palate, and cardiac and dental anomalies. There is considerable overlap between the CXN, NHS and OFCD phenotypes, emphasising genetic heterogeneity on the X-chromosome for these related developmental disease and raising the question as to whether mutations in *BCOR* could be disease causative in CXN and NHS. Currently only one family with congenital cataract has been mapped to Xp22.13-22.2 (Francis *et al.*, 2002). The possibility that a double cross over occurring with the CXN family may have led to the wrong assignment of the CXN disease locus to Xp22.13-22.2 in the original study. However, the new haplotypes created in this thesis using novel microsatellites, confirmed segregation with Xp22.13-22.2. In addition, the typing of microsatellite markers spanning the X chromosome resulted in crossovers being observed between disease in the CXN family and all markers residing on Xp11.4, in particular microsatellite markers DXS8015 and DXS993. Microsatellite marker DXS8015 resides 0.03 Mb upstream of the *BCOR* gene, whilst DXS993 resides 1.09 Mb downstream of the *BCOR* gene. Therefore disease segregation with Xp11.4 was not identified from genotyping microsatellite markers in the CXN family in this study.

Since the families with Nance-Horan syndrome in this study are of a small size, haplotype analysis was not performed in order to confirm segregation of disease with

markers on Xp22.13. It has not been possible therefore, to determine whether NHS is heterogenous by mapping NHS families to alternative loci. The possibility that some of the NHS families in this study may map elsewhere on the X chromosome and could harbour mutations within the *BCOR* gene cannot be ruled out.

The causative gene(s) for CXN and NHS remained to be identified.

- Chapter Four -

Identification of the gene for Nance-Horan Syndrome (NHS)

4.1 Introduction

Following refinement of the CXN locus and the exclusion of 7 candidate genes (*RAI2*, *TL1*, *RBBP7*, *CX43-L*, *PPEF1*, *RS1*, *and CDKL5*) for CXN and/or NHS, a comprehensive genomic survey of the region of overlap between the NHS and CXN loci was performed (figure 3.11).

4.2 Results

4.2.1 *Novel candidate gene 1* (*NCG1*), the Nance-Horan syndrome gene

Using comparative genomics and bioinformatic tools (section 2.7), novel genes were sought and partially characterised to identify potential candidates for disease. As a result, a novel gene distal to *SCML1* and proximal to microsatellite marker DXS1195 (AFM207ZD6) was identified and found to be highly conserved in mouse. The novel gene (*NCG1*) was predicted to contain at least 9 coding exons (table 4.1, figure 4.1) spanning at least 65 Kb of genomic sequence and coding for a 1,471 amino acid protein.

Exon	Size (bp)	Exon-intron junction	
		Acceptor	Donor
1			GTTGGCAG**gt**cagcat
2	152	ccctgc**ag**GTGCCAAC	AAGGGCCA**gt**ggcaga
3	153	tcttgc**ag**CCGTCTCC	GCGCAGAG**gt**gacaga
4	134	cttctc**ag**AACACCGG	TTTTAACG**gt**aagttt
5	193	gtttgc**ag**TCCCATCC	TGTTACTG**gt**atcgtt
6	132	ggttgc**ag**GAGTTGGC	AGAAATAG**gt**gtgata
7	2982	tgccct**ag**ATTCTGAT	ATTGAAAG**gt**cagtca
8	127	atttta**ag**AATCATCA	ATTCACAG**gt**gaggca
9		tctcaa**ag**ATCCAAGA	GCAAATAAACGTGACT

Table 4.1 Human *NCG1* genomic structure

The human gene prediction was supported by numerous EST hits (NCBI Unigene cluster Hs.444940) expressed in multiple tissues including fetal eye, brain, kidney, muscle, heart, liver, lung and colon. NIX analysis of the predicted gene identified a mRNA hit (accession number AK026164) joining together part of exon 7 with exon 8 and the terminal exon (exon 9). In addition, part of this cDNA sequence had high homology to the myosin gene family.

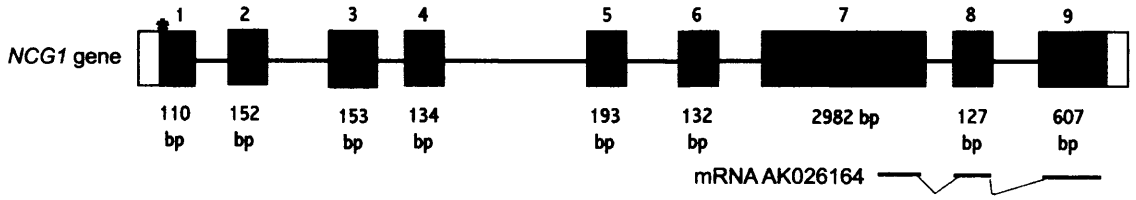


Figure 4.1. Genomic structure of the *NCG1* gene.

Boxes represent exons, blackened boxes indicate the open reading frame (ORF), with sizes of the coding portions given below. Green asterisk denotes the methionine start codon.

Based on the position of *NCG1* within all three disease loci (CXN, NHS and Xcat, figure 4.2), its expression in relevant tissues and its potential similarity to myosins, all coding exons and splice sites were screened in an affected male from the CXN family and all the NHS families. Primers were designed (section 2.2.2, table 4.2) to PCR amplify (section 2.2.1) all 9 predicted coding exons and splice sites.

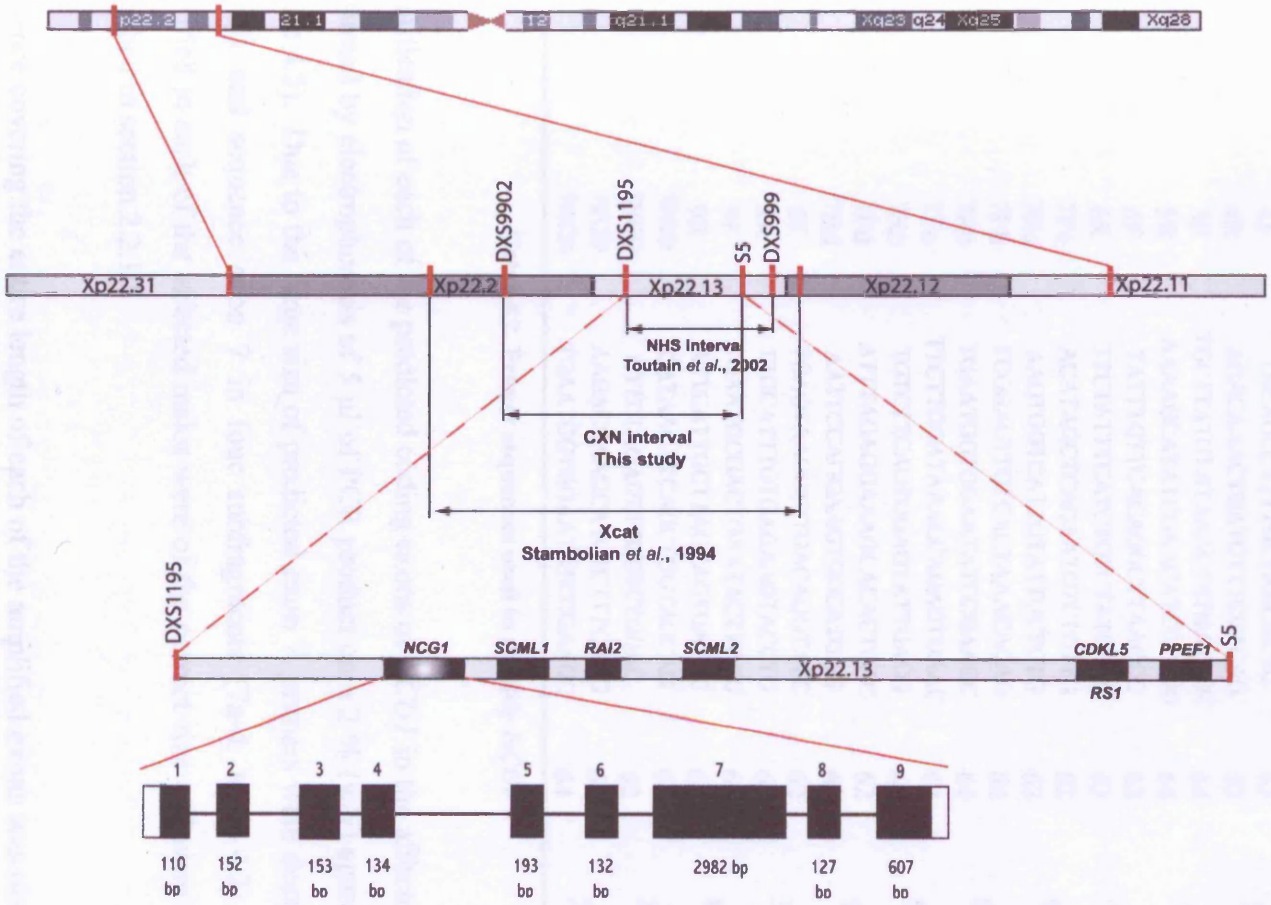


Figure 4.2 Identification of *NCG1* within the region of overlap between disease loci CXN, NHS, and Xcat

Relative position of the *NCG1* gene prediction with relation to other genes on Xp22.13

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
NCG1	1F	TTAACTTGATTGCCTGCAACTG	62	345
	1R	TTCAGAGACTTTGAATGAGATAAG	62	
	2F	TTGTGGAGTTTCCATCATAGG	60	281
	2R	TTGCAGTATTTCGAAAACTAGG	60	
	3F	GACTTTACTCCTGTCCTCTTC	62	519
	3R	ATGCCTGAGCATTTGCATATTG	62	
	4F	TACAGCCTTTTGCTAACACTC	62	281
	4R	AGACAAACTGATGTTTCCTCAG	62	
	5F	TGCTTATGTATAAACTATGATGGC	64	397
	5R	AAAAGCATATCGAACATCCTCTG	64	
	6F	TATTTGTTCACAGGCTTAACCC	62	327
	6R	TTCTATTTCATCTCCCTATCATC	62	
	7Fa	ACATAGCTCAGAATGTCTGTTG	62	901
	7Ra	AAGTGGTCATTGTATTGCTCTG	62	
	7Fb	TCGGAGTTGTCACTAAACACAG	64	814
	7Rb	TGAATGGTGGAATATCCGAAGC	64	
	7Fc	TTCTTGGATAAACCAGAGTGAAC	64	857
	7Rc	TGTCTTCAGTGGAGTATTGACG	64	
	7Fd	ATTTAGAGGAAAGCACACTCAC	62	790
	7Rd	AATTCCAGGAAGTGCCATGAG	62	
	8F	TGAGTAAGCTTTGACAGGTATC	62	364
	8R	TTGCATTTGTGAGAAGTACCTG	62	
	9F	TGAATGCGACTGAATACTTCAG	62	807
	9R	ATTCATTGCTTAGCAGTGATGC	62	
	7FBD	CATAACATCAGCTGGTAGCAG	62	207
	7RBD	GTTCTAGAGTTTCGTCTGGAC	62	
	7FCD	AAGACCAAGCATCTCTTTCAGG	64	299
	7RCD	TGAATGGTGGAATATCCGAAGC	64	

Table 4.2 Primer sequences used to amplify *NCG1*

Amplification of each of the predicted coding exons of *NCG1* in the affected males, was confirmed by electrophoresis of 5 µl of PCR product on a 2 % (w/v) agarose check gel (figure 4.3). Due to the large size of predicted exon 7, primers were designed so as to amplify and sequence exon 7 in four subfragments (7a-d, table 4.2). All exons amplified in each of the affected males were of the correct size and were sequenced as described in section 2.2.10.

Sequence covering the entire length of each of the amplified exons was obtained. From bioinformatic analysis of genomic sequence on the database, no stop codon was identified in the 3' exon (exon 9) of the *NCG1* gene prediction. Upon sequencing exon 9 in the affected males and a control DNA sample it became evident that the stop codon was omitted from the current genomic sequence available (June 2003).

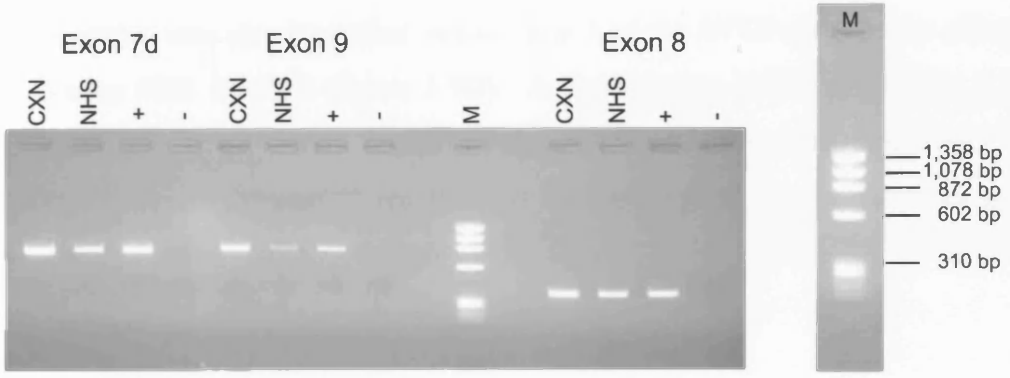


Figure 4.3 Examples of PCR amplicons for *NCG1* gene exons from an affected male from the CXN and NHS family 1, resolved on a 2 % (w/v) agarose gel.

CXN = affected male from the CXN family, NHS = affected male from NHS family 1; +, control DNA; -, no DNA control. Anticipated sizes of individual exons are: 7d (790 bp), 8 (364 bp), 9 (807 bp). Molecular weight marker (M) was run alongside to confirm the size of PCR products, details of markers are shown on the right hand panel

No mutations were identified in any of the exons for NHS families 2 and 4, or the CXN family. However, sequencing of fragment d of exon 7 of the *NCG1* gene in NHS family 1 (affected male IV:3, figure 3.3B) resulted in the identification of a 2 bp deletion (delTG) at position 3,261-3,262 bp of the *NCG1* Genscan gene prediction. Deletion of the two bases (TG) resulted in a frameshift leading to the addition of 15 novel amino acids followed by a premature stop codon (figure 4.4).

```
Wildtype:       caagtgcgtacagagactgagcctattccagaaaacacgccaacc
                Q  V  R  T  E  T  E  P  I  P  E  N  T  P  T
NHS family 1:   caagtgcgtacagagactgagcctattccagaaaacacgccaacc
                Q  V  R  T  E  T  E  P  I  P  E  N  T  P  T

                        **
Wildtype:       aaaaactgtgctttcccacagaaggatttcagagggtctctgct
                K  N  C  A  F  P  T  E  G  F  Q  R  V  S  A
NHS family 1:   aaaaactgctttcccacagaaggatttcagagggtctctgctgc
                K  N  C  F  S  H  R  R  I  S  E  G  L  C  C

Wildtype:       gcccgcccaaatgatttggatggtaaaataatacaatatggacct
                A  R  P  N  D  L  D  G  K  I  I  Q  Y  G  P
NHS family 1:   ccgcccaaatga
                P  P  K  *
```

Figure 4.4 Sequence of NCG1 in an affected male from NHS family 1 compared with wildtype sequence

A 2 bp (TG) deletion identified in an affected male from NHS family 1 results in a frameshift leading to an additional 15 novel amino acids followed by a premature stop codon.

A second mutation was also identified within exon 7 of the *NCG1* gene in the affected male (II:1) from NHS family 3 (figure 3.3D). A 1bp deletion (delA) at position 2,210 bp of the Genscan gene prediction for *NCG1* was identified in this family member. The 1 bp deletion causes a frameshift resulting in the addition of 10 novel amino acids followed by a premature stop codon (figure 4.5).

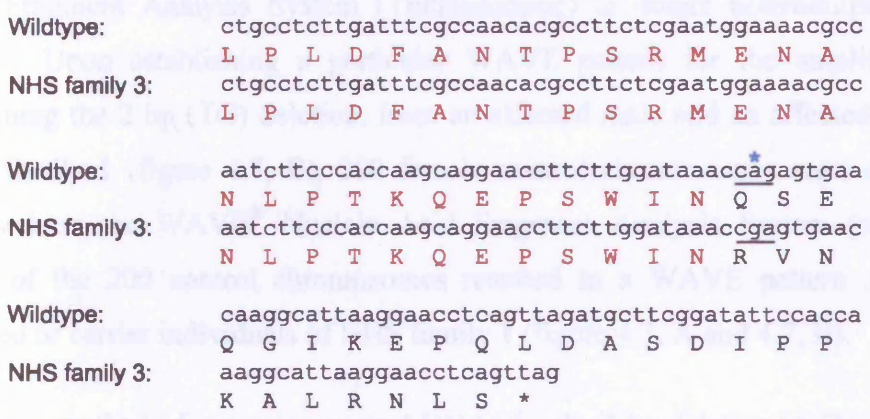


Figure 4.5 Sequence of NCG1 in the affected male from NHS family 3 compared with wildtype sequence

A 1 bp (A) deletion identified in the affected male from NHS family 3 results in a frameshift leading to an additional 10 novel amino acids followed by a premature stop codon.

Upon identifying deletion mutations within exon 7 of the *NCG1* gene in two families with NHS (NHS families 1 and 3), segregation of the 2 bp deletion (delTG) with disease in family 1 was confirmed. It was not possible to test segregation of the 1 bp deletion (delA) identified in NHS family 3 as only one DNA sample for the affected son (individual II:1, figure 3.3D) was available. Exon 7 (fragment d) was amplified (section 2.2.1) in each member of NHS family 1 for whom DNA was available.

Sequencing of exon 7d in each member of NHS family 1 confirmed segregation of the 2 bp deletion (delTG) with disease in this family (figure 4.6). Identification of two deletions both causing frameshifts, which are predicted to result in the addition of novel amino acids prior to a premature stop codon strongly implied that the *NCG1* gene prediction is the NHS gene. To confirm the original 2 bp deletion observed in NHS

family 1 was disease causing, 200 control chromosomes were screened for this sequence change.

Primers were designed (section 2.2.2; table 4.2, primers 7FBD and 7RBD) to amplify 207 bp across the 2 bp (TG) deletion in NHS family 1. PCR amplification (section 2.2.1) across the 2 bp (TG) deletion was carried out on a carrier female and an affected male of NHS family 1. The amplicons were then analysed using the WAVE® Nucleic Acid Fragment Analysis System (Transgenomic) to detect heterodulplexes (section 2.2.8). Upon establishing a particular WAVE pattern for the amplified fragment containing the 2 bp (TG) deletion, from an affected male and an affected female, from NHS family 1 (figure 4.7, B), 200 female control chromosomes were amplified and analysed on the WAVE® Nucleic Acid Fragment Analysis System (section 2.2.8). None of the 200 control chromosomes resulted in a WAVE pattern similar to the affected or carrier individuals of NHS family 1 (figure 4.7, A and 4.7, B).

The same method of screening control DNAs for the 2 bp deletion (delTG) identified in NHS family 1, was carried out for the 1 bp deletion (delA) identified in NHS family 3. Primers were designed (section 2.2.2; table 4.2, primers 7FCD and 7RCD) to amplify 299 bp across the 1 bp (delA) in NHS family 3 (figure 4.8). The primers were then used to amplify across the 1 bp (A) deletion by PCR (section 2.2.1) in the affected male of NHS family 3 (DNA was not available for the carrier female in NHS family 3). Once a particular WAVE pattern for the amplified fragment containing the 1 bp (A) deletion was established, 200 control chromosomes were amplified (figure 4.8) and analysed on the WAVE® Nucleic Acid Fragment Analysis System (section 2.2.8). None of the 200 control chromosomes resulted in a WAVE pattern similar to the affected male of NHS family 3.

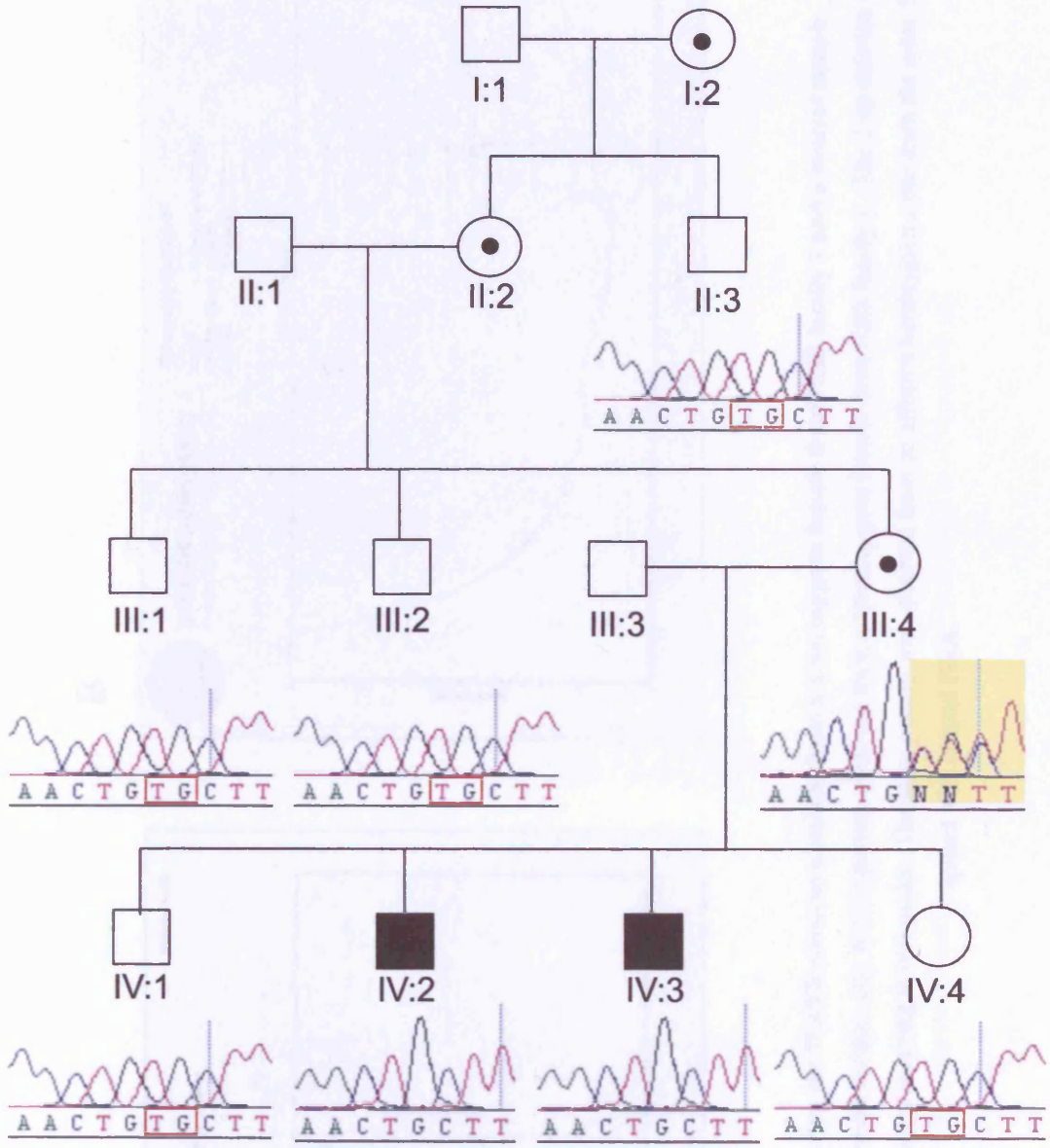
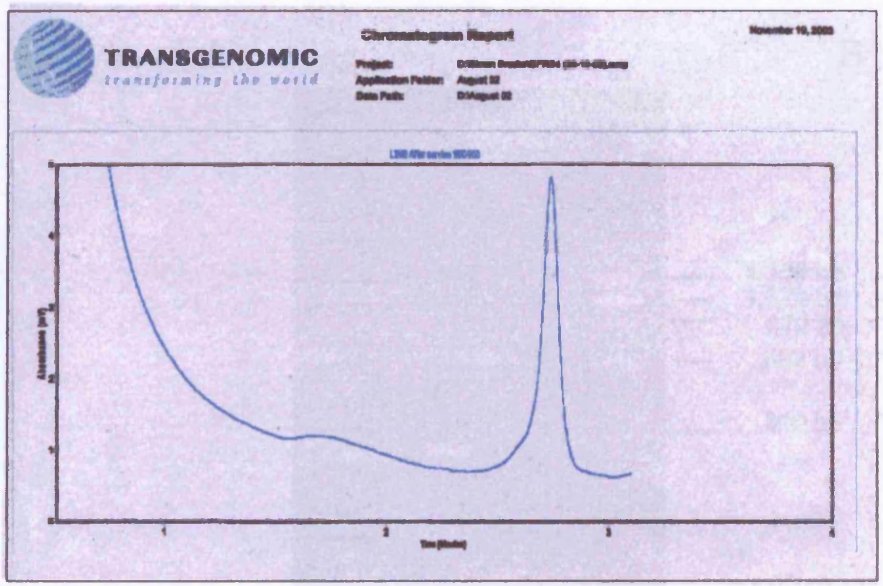


Figure 4.6 Segregation of a 2 bp deletion (delTG) with disease in NHS family 1.

Sequence chromatograms for two affected boys, one carrier female, and 5 unaffected individuals demonstrates segregation of the 2 bp deletion with disease in NHS family 1. The two bases, TG, deleted in the two affected boys and in the carrier female is boxed red in the unaffected individuals. No DNA sample was available for individual III:3.

A



B

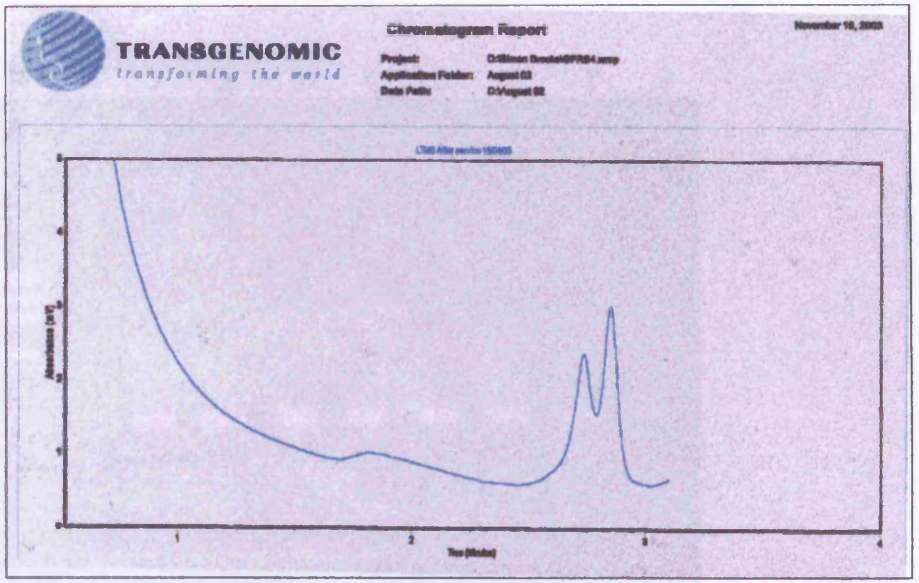


Figure 4.7 Chromatograms showing the WAVE patterns obtained from a heterozygous female from NHS family 1 and a control sample.

(A) WAVE pattern observed from a control female sample. (B) WAVE pattern observed from a heterozygous female from NHS family 1. The 2 bp deletion results in a heteroduplex, which elutes prior to the homoduplex resulting in two peaks. The same pattern was observed from an affected hemizygous male from the same family when spiked with control DNA.

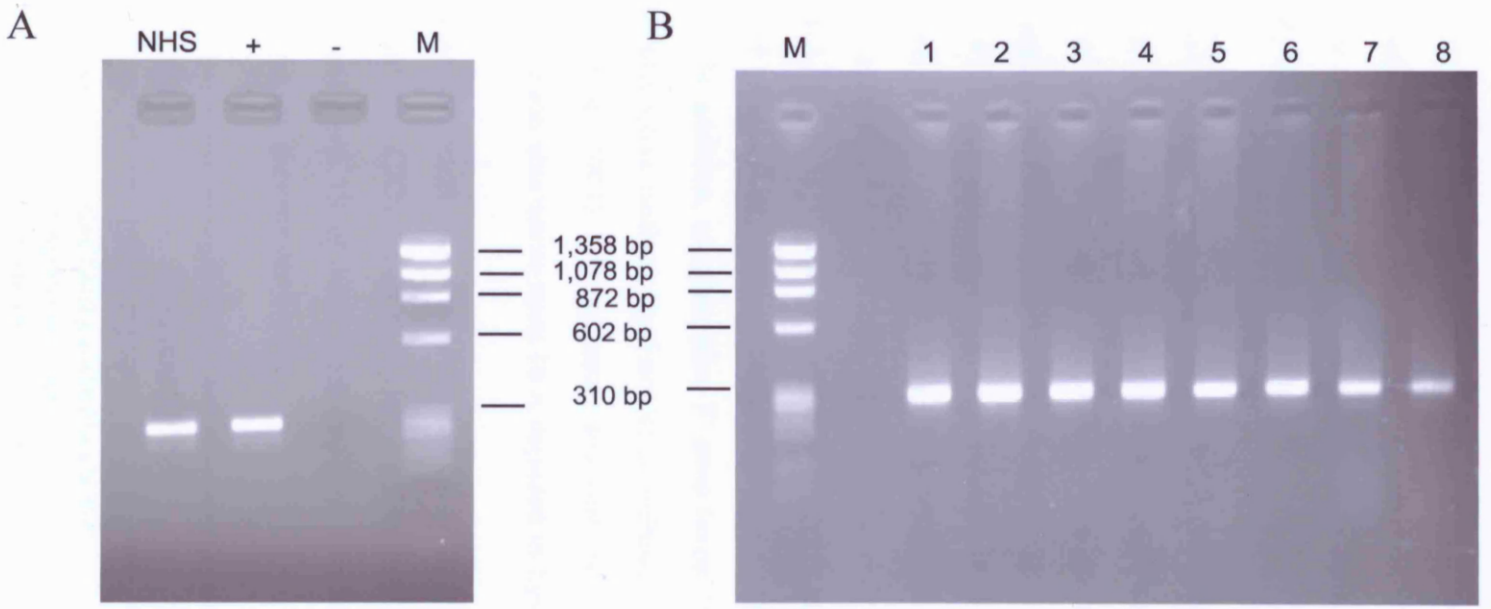


Figure 4.8 Amplification across the 1 bp deletion identified in NHS family 3, and amplification of the same region in control DNAs.

(A) Amplification across the 1 bp deletion identified in NHS family 1, in the affected male. +, control DNA; M, marker. (B) Amplification of the same fragment in 8 control DNA samples; M, marker. All amplicons were resolved on a 2 % (w/v) agarose gel.

4.2.2 Further characterisation of the genomic structure of *NCG1*, the Nance-Horan syndrome gene

Since *NCG1* was evidently the Nance-Horan syndrome gene as two different frameshifting deletions had been identified in Nance-Horan syndrome families 1 and 3, further characterisation of the gene was required to determine the full genomic structure and identify further mutations in the two remaining NHS families and the CXN family. *NCG1* is here after referred to as the *NHS* gene.

During further analysis of the gene structure for *NHS*, Burdon and colleagues reported the finding of a novel gene in which truncation mutations cause the pleiotropic effects of Nance-Horan syndrome (Burdon *et al.*, 2003). The gene was predicted by gene prediction programmes and the transcript was further characterised through northern blot analysis and RT-PCR. The *NHS* gene they describe is the same gene as the novel gene identified here as the cause of NHS (*NCG1*). Burdon *et al.*, reported the *NHS* gene to contain 9 exons of which 8 were coding. Exon 1a of *NHS* was found to be ~350 kb upstream of exon 2. Exons 2, 3, 4, 5, 6, 7 and 8 of the *NHS* gene proved to be exons 3, 4, 5, 6, 7, 8 and 9 (respectively) of the *NCG1/NHS* gene prediction described in this thesis (figure 4.1). In addition, an alternative 5' exon (exon 1b) is reported in the *NHS* gene, forming an alternative isoform (isoform b) containing exons 1b and 4 – 8 of the *NHS* gene (Burdon *et al.*, 2003). The genomic structure and isoforms of the *NHS* gene, with the new exon 1a and alternative exon 1b is depicted in figure 4.9.

As a result of these additional data, exons 1a and 1b of the *NHS* gene were amplified and sequenced in both the CXN family and NHS families 2 and 4. Primer sequences for the new exon 1a and exon 1b of *NHS* were designed (section 2.2.2, table 4.3) from sequence available on the genome database.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
NHS	1aF	TATCCGGACTGCCAGATCGC	64	815
	1aR	GAGTAGTAAGGTGCAAGCTGC	64	
	1aFB	AGGGACCTGGACGAGGTC	62	
	1bF	AATGAAGCCTCTGGTGGTGG	60	446
	1bR	ATGCCCTGTTGACTCAGAGG	62	

Table 4.3 Primer sequences used to amplify exons 1a and 1b of the *NHS* gene

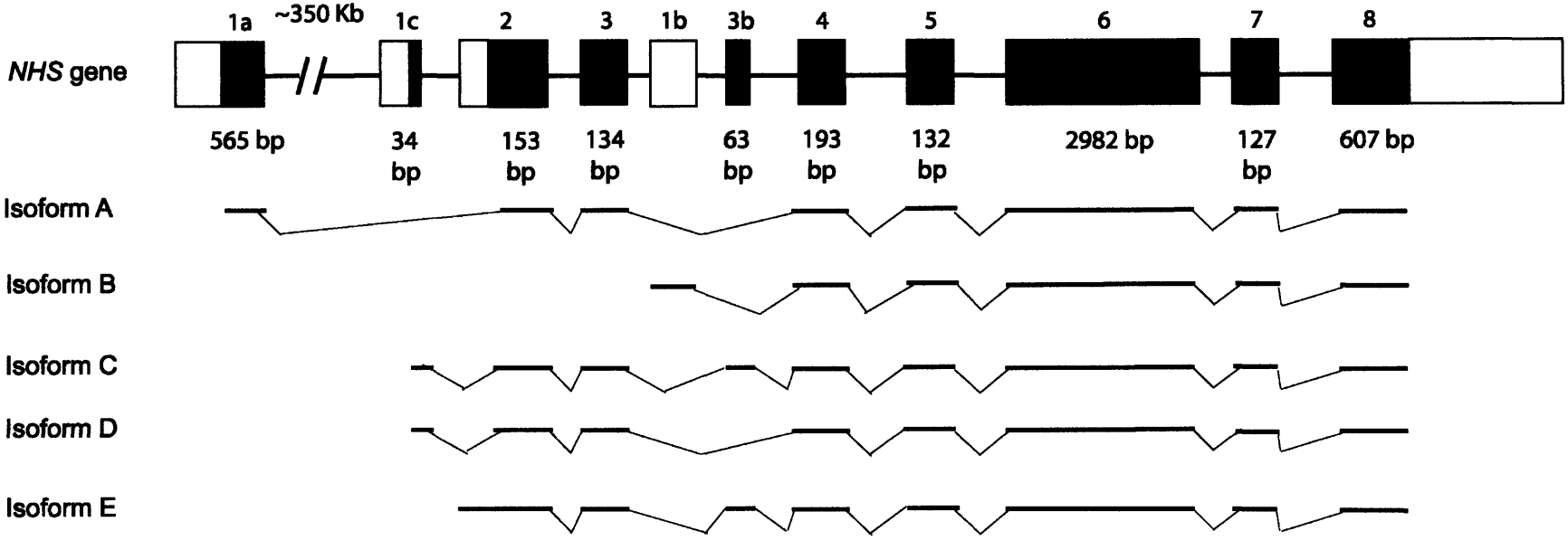


Figure 4.9 Genomic structure of the *NHS* gene.

A number of alternative isoforms have been identified on the database. Currently, there are five isoforms (isoforms A – E). Blackened squares indicate the open reading frame (ORF), with sizes of coding portions given below. Exons 2, 3, 4, 5, 6, 7 and 8 of the *NHS* gene proved to be exons 3, 4, 5, 6, 7, 8 and 9 (respectively) of the *NCG1* gene prediction described prior to the publication of the Burdon and colleagues paper (Burdon *et al.*, 2003).

Using the primers designed to the new exon 1a and 1b of the *NHS* gene, each exon was PCR amplified (section 2.2.1, figure 4.10) from an affected male from NHS families 2 and 4, and from the CXN family. Amplicons were then sequenced directly (section 2.2.10) in the hope of identifying a mutation in one of these exons in each affected male.

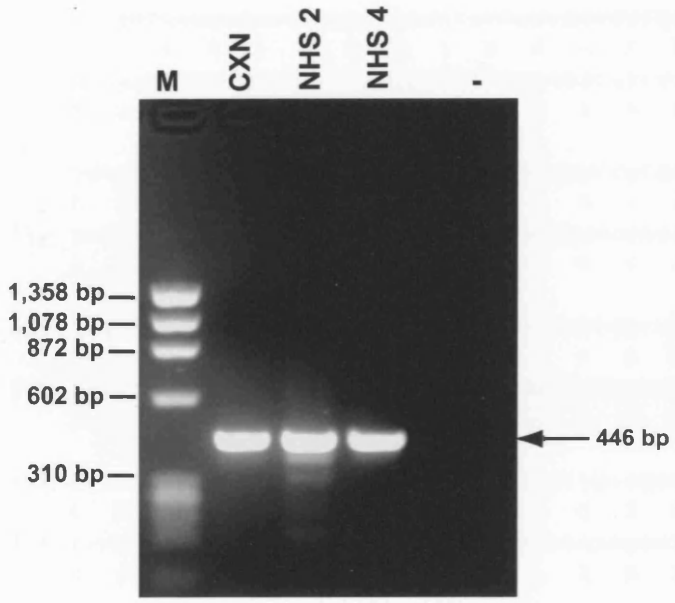


Figure 4.10 PCR amplification of *NHS* exon 1b in an affected male from the CXN family, and NHS families 2 and 4, resolved on 2 % (w/v) agarose gel

M, marker; -, no DNA control

No sequence alterations were identified in the CXN family. However, a 1 bp deletion (400delC) was identified in exon 1a of NHS family 2, and is predicted to cause a frameshift resulting in a premature stop codon after the addition of 61 novel amino acids (figure 4.11), and a severely truncated protein product.

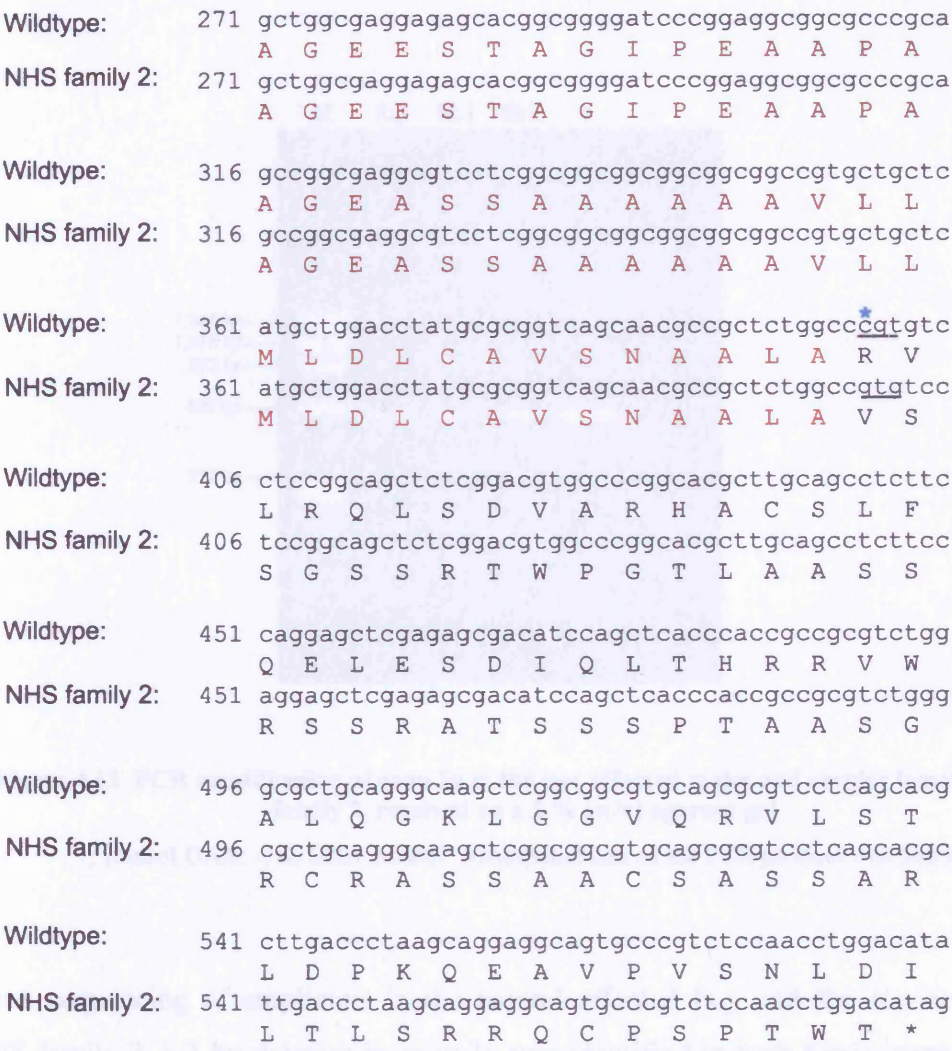


Figure 4.11 Sequence of NHS in an affected male from NHS family 2 compared with wildtype sequence

A 1 bp (C) deletion (400delC) identified in exon 1a of an affected male from NHS family 2, results in a frameshift leading to an additional 61 novel amino acids followed by a premature stop codon.

Segregation of the 1 bp deletion (del400C) in NHS family 2 was confirmed by PCR amplification (section 2.2.1) and sequencing (section 2.2.10) exon 1a in members of the NHS family 2 for whom DNA was available. A newly designed primer NHS 1aFB (table 4.3) was designed (section 2.2.2) for amplification across the 1 bp deletion in NHS family 2 (figure 4.12).

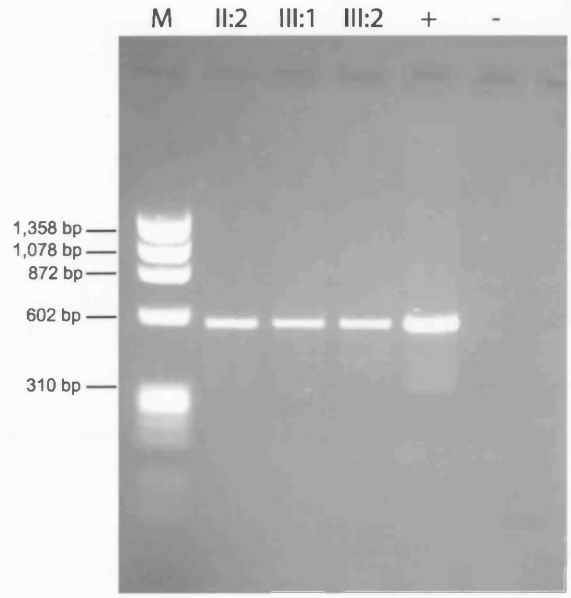


Figure 4.12 PCR amplification of exon 1a in the two affected males and carrier female of NHS family 2, resolved on a 2 % (w/v) agarose gel

+, control DNA; -, no DNA control. Anticipated size of the PCR product was 560 bp.

Upon sequencing of amplicons in the second affected boy and the carrier female of NHS family 2, a 1 bp deletion in exon 1a was identified in both family members. The female, as expected, was found to be heterozygous for the mutation (figure 4.13).

Using the same primers (1aFB with 1aR, size 560 bp) to amplify across the 1 bp deletion (400delC) in individuals of NHS family 2, 200 control chromosomes were PCR amplified (section 2.2.1) and sequenced (2.2.10) to determine whether the 1 bp deletion was a polymorphic variant. No sequence changes were identified in any of the 200 control chromosomes.

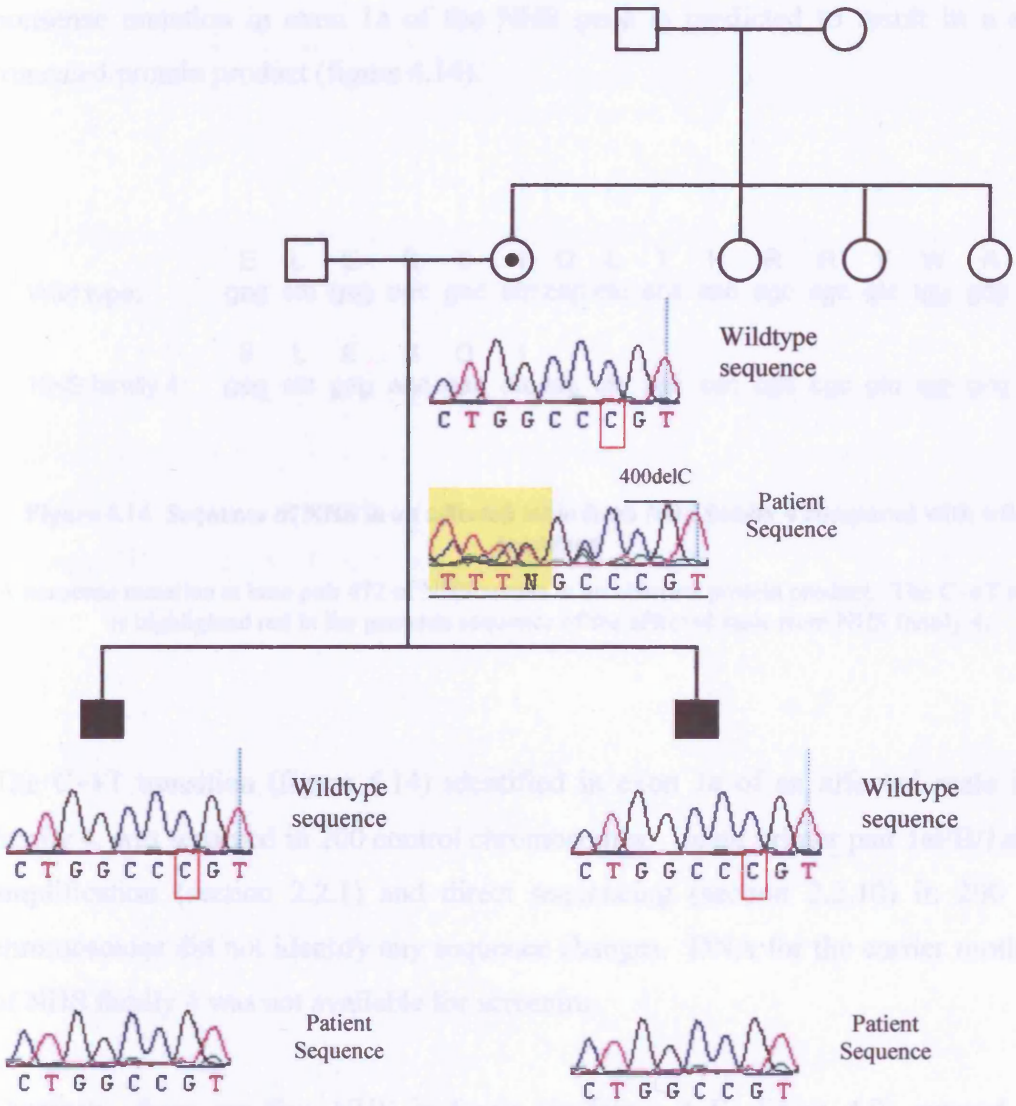


Figure 4.13 Segregation of a 1 bp deletion (delA) in exon 1a of *NHS* with disease in NHS family 2.
Sequence chromatograms for two affected boys and one carrier female demonstrates segregation of the 1 bp deletion (400delC) with disease in NHS family 2. The base deleted in the two affected boys and in the carrier female is boxed in red in the wildtype sequence. No unaffected individuals in this family were available for this study. The wild type sequence is given above each chromatogram from the two affected males and carrier females for comparison.

Screening of exon 1a in an affected male from NHS family 4 revealed a nonsense mutation. A C→T transition (figure 4.14) at the first position of codon 158 (472C→T) alters the sequence from a glutamine (Q) to a stop (figure 4.14). The introduction of a nonsense mutation in exon 1a of the NHS gene is predicted to result in a severely truncated protein product (figure 4.14).

```
                E  L  E  S  D  I  Q  L  T  H  R  R  V  W  A
Wild type:      gag ctc gag agc gac atc cag ctc acc cac cgc cgc gtc tgg gcg

                E  L  E  S  D  I  .
NHS family 4:   gag ctc gag agc gac atc tag ctc acc cac cgc cgc gtc tgg gcg
```

Figure 4.14 Sequence of NHS in an affected male from NHS family 4 compared with wildtype sequence

A nonsense mutation at base pair 472 of NHS results in an aberrant protein product. The C→T transition is highlighted red in the genomic sequence of the affected male from NHS family 4.

The C→T transition (figure 4.14) identified in exon 1a of an affected male in NHS family 4, was screened in 200 control chromosomes. Using primer pair 1aFB/1aR, PCR amplification (section 2.2.1) and direct sequencing (section 2.2.10) in 200 control chromosomes did not identify any sequence changes. DNA for the carrier mother (I:1) of NHS family 4 was not available for screening.

Currently, there are five NHS isoforms (isoforms A-E, figure 4.9) entered on the database. Isoforms C-E contain alternative 5' exons, and a cryptic exon 3b, which is spliced in, in isoforms C and E (figure 4.9). Primers were designed (section 2.2.2, table 4.4) to PCR amplify (section 2.2.1) and sequence (section 2.2.10) these additional *NHS* exons for screening in an affected male from the CXN family.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
NHS	1cF	AAGTCCCACTGCAACCTAAAC	62	
	1cR	AAGTGGCTACAGCACAAGCC	62	489
	3bF	ATACACTGTGTTGTGTGCACG	62	
	3bR	TCTGGACAGAGTGGGATAGG	62	171

Table 4.4 Primer sequences used to amplify altenative 5' exons and exon 3b identified in isoforms C-E of the *NHS* gene.

No sequence variants were identified in the alternative 5' exon or exon 3b of *NHS* isoforms C-E, in an affected male from the CXN family.

Figure 4.15 summarises all mutations identified in the *NHS* gene in all four NHS families screened in this study. In addition, table 4.5 summarises all mutations identified in the *NHS* gene from this thesis, along with those reported by Burdon and colleagues (2003) and by Ramprasad and colleagues (2005).

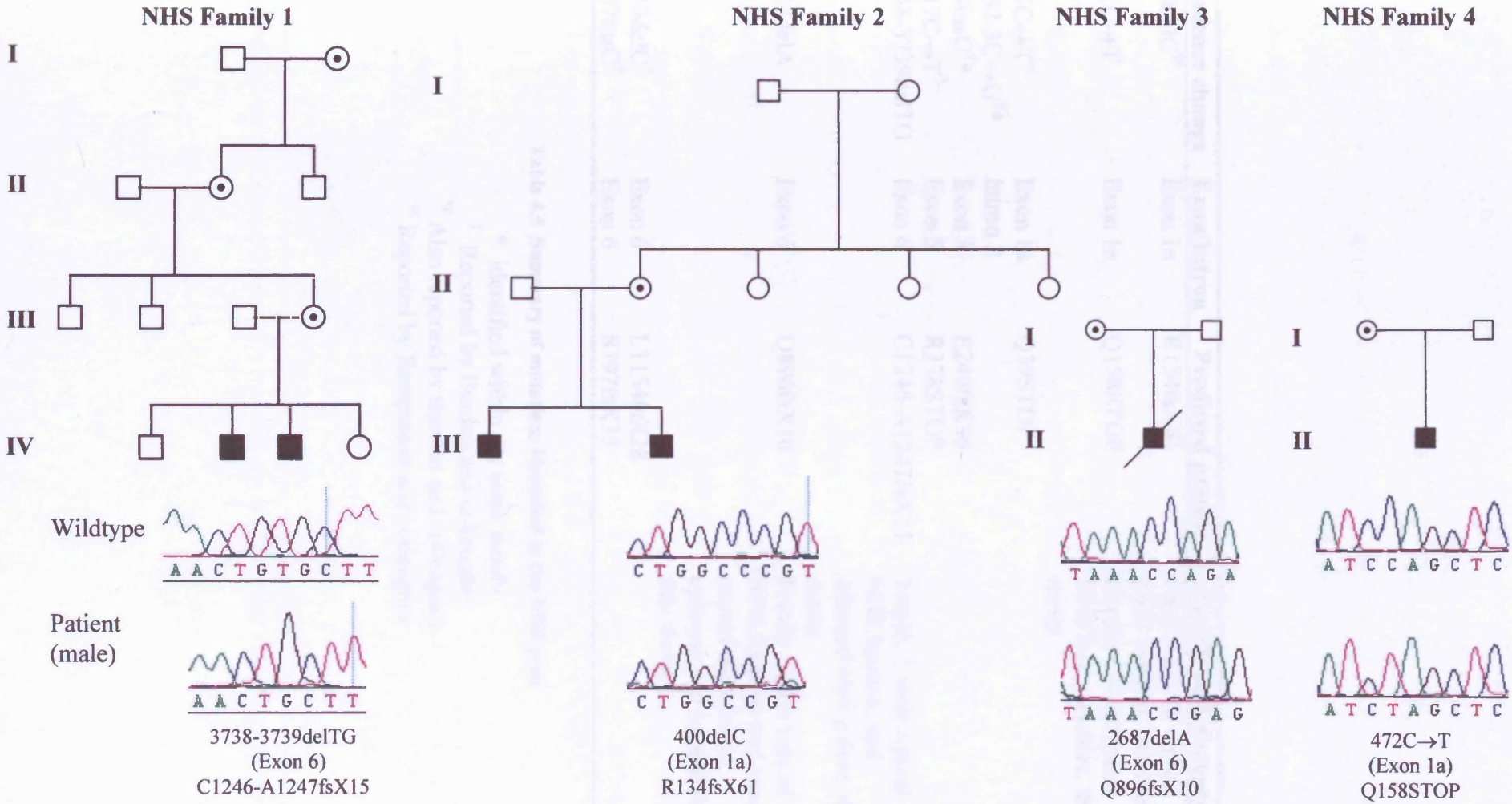


Figure 4.15 NHS families analysed in this study and corresponding sequence chromatograms highlighting mutations identified in the *NHS* gene

Sequence change	Exon/Intron	Predicted protein	Family and phenotype
400delC[Ψ]	Exon 1a	R134fsX61	Family 2 with typical NHS features, this thesis
472C→T	Exon 1a	Q158STOP	Family 4 with typical NHS lens opacities, this thesis
115C→T[∝]	Exon 1a	Q39STOP	
IVS2-3C→G[†]*	Intron 2		
718insG[†]*	Exon 3	E240fsX36	
1117C→T[†]	Exon 5	R378STOP	
3738-3739delTG	Exon 6	C1246-A1247fsX15	Family 1 with typical NHS features and bilateral cleft palate, this thesis
2687delA	Exon 6	Q896fsX10	Family 3 with typical NHS features and severe mental retardation, epilepsy and hypotonia, this thesis
3459delC[†]	Exon 6	L1154fsX28	
2387insC[†]	Exon 6	S797fsX35	

Table 4.5 Summary of mutations identified in the *NHS* gene

* identified within the same family
[†] Reported by Burdon and colleagues
[Ψ] Also reported by Burdon and colleagues
[∝] Reported by Ramprasad and colleagues

4.2.2.1 Expression of *NCG1/NHS*

4.2.2.1.1 AK026164

Initial characterisation of the genomic structure of *NCG1* was based on gene prediction programmes and comparative genomics, supported by ESTs and an mRNA (AK026164) aligning with predicted exons of the gene. Interestingly, on further analysis mRNA AK026164 (from human renal epithelial cells) was found to be a chimera. AK026164 (1,741 nucleotides in size) contained sequence from exons 7, 8 and 9 of the *NCG1* gene prediction (first 1,016 nucleotides), whilst the remaining 725 nucleotides were found to match non-muscle myosin light chain alkali, on chromosome 12. Therefore, the novel gene identified proved not to share homology with the myosin protein family.

4.2.2.1.2 Human fetal cDNA panel

As NHS is a developmental syndrome, the expression of the *NHS* gene in human fetal tissue was investigated. A human fetal multiple tissue cDNA (MTC) panel was obtained from BD Biosciences (UK), and primers were designed (section 2.2.2, table 4.6) to amplify from exons 6-7 of the *NHS* gene, which is present in all isoforms. Amplification spanning intron 6 of the *NHS* gene allowed for contamination of the human fetal cDNA panel with genomic DNA to be tested. Amplification of *NHS* from cDNA was predicted to result in a PCR fragment of 268 bp, whereas amplification of *NHS* from contaminating genomic DNA should result in a product of 590 bp.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
	NHScDNA6F	TGCCTGGTACTATCAGCTATG	64	268
	NHScDNA7R	TCCTCAGTTGTTCGAGGTTTG	62	
	Mouse *Nhs* 6F	ACTGCAACCTCAGCTAGCAG	62	207
	Mouse *Nhs* 6R	TGCTGTACAGTGCCATCTGG	62	

Table 4.6 Primer sequences used to amplify the *NHS* gene in a human fetal cDNA panel and mouse tooth cDNA.

Expression of the *NHS* gene in each of the 8 human fetal tissues (spleen, brain, liver, kidney, heart, lung, thymus, and skeletal muscle) was determined by attempting to PCR amplify exons 6-7 of the *NHS* gene using the newly designed primers mentioned above (table 4.6). A 268 bp fragment was amplified from human fetal brain, thymus, lung, and kidney (figure 4.16), demonstrating expression of the *NHS* gene in these tissues. Further, amplification of the *NHS* gene in these tissues was not a result of DNA contamination, as a 590 bp product was not observed.

Since tooth abnormalities are a typical clinical feature of Nance-Horan syndrome, expression of the mouse *NHS* orthologue in mouse tooth cDNA at postnatal day zero, P0, and postnatal day two, P2, was investigated. RNA was extracted from mouse teeth at P0 and P2 (section 2.2.9) and reverse transcribed (section 2.2.3). Primers were designed (section 2.2.2, table 4.6) to amplify a region within exon 6 of the mouse *NHS* orthologue (see chapter 5, section 5.2.1). PCR amplification (section 2.2.1) of the mouse *Nhs* gene from mouse tooth cDNA was successful at P0, but not P2 (figure 4.16).

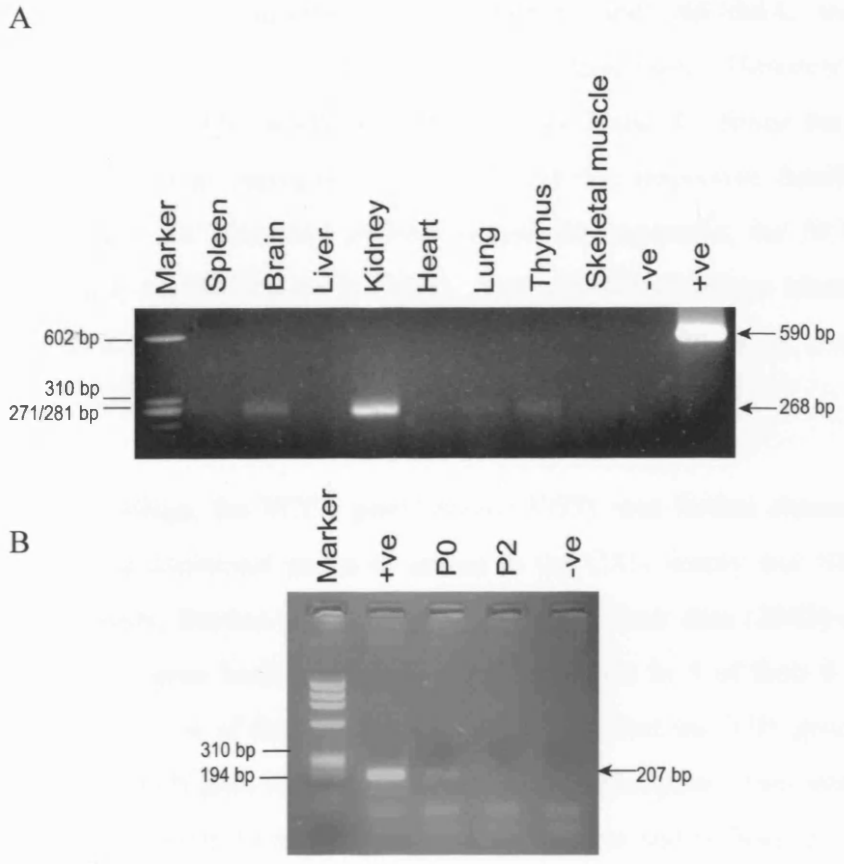


Figure 4.16 PCR from human fetal cDNA (panel A) and mouse tooth cDNA (panel B) at P0 and P2

Primers were designed to amplify exon 6 – 8 of the *NCG1* gene; +ve, genomic DNA control; -ve, no DNA control. PCR products were resolved on a 2 % (w/v) agarose gel.

4.2.3 Discussion

Following the exclusion of all characterised candidate genes residing in the disease interval for Nance-Horan syndrome, novel candidate genes were sought in the region of overlap between the NHS and CXN disease loci. As a result a Genscan gene prediction, here named *NCG1*, was identified and screened in the four NHS families and the CXN family. Two deletion mutations, 3738-3739delTG and 2687delA, were identified within exon 6 of *NCG1* in NHS families 1 and 3 respectively. However, no mutation was identified in the CXN family or NHS families 2 and 4. Since the two deletion mutations were found to segregate with disease in the respective families and since these changes were not identified in 200 control chromosomes, the *NCG1* gene was clearly the disease causative gene for NHS. Further, the mutations identified in NHS families 1 and 3 disrupted the open reading frame resulting in an aberrant protein product.

Based on these findings, the *NCG1* gene (named *NHS*), was further characterised in the hope of identifying additional exons to screen in the CXN family and NHS families 2 and 4. Subsequently, Burdon and colleagues published their data (2003) detailing their findings of a novel gene harbouring truncating mutations in 5 of their 6 families with NHS. Closer inspection of their findings demonstrated that the *NHS* gene identified in this thesis was the *NHS* gene reported by Burdon and colleagues. Two additional novel, alternative 5' exons (exon 1a and 1b) reported by Burdon and colleagues (2003), which had not been screened in the CXN family and NHS families 2 and 4, were amplified and sequenced in an affected male from each of these families (Burdon *et al.*, 2003).

A 1 bp deletion (400delC) was identified in exon 1a of NHS family 2, whilst a nonsense mutation (472C→T) was found in exon 1a of NHS family 4. No sequence alterations were identified in either exon 1a or exon 1b of the CXN family. The sequence changes identified in NHS families 2 and 4 (400delC and 472C→T, respectively) were confirmed as disease causative in these families by screening 200 control chromosomes for these sequence variants. Neither change was observed in any of the 200 control chromosomes screened.

Burdon and colleagues identified truncating mutations in 5 independent NHS families. Within the index case of one family, a truncating mutation and a one base pair

substitution at the 3' acceptor splice site were identified in exon 3. This female with NHS is reported to display an unusually severe ocular phenotype. One of the NHS families reported by Burdon and colleagues appears to be the same family as NHS family 2, reported in this study. Their reporting of a 1 base pair deletion (ΔC) at the same position in exon 1a (base pair 400) and personal communications via a common clinical colleague, confirmed this.

A more recent paper by Ramprasad and colleagues (2005), reports the finding of a nonsense mutation (115C→T) in exon 1a of the *NHS* gene. The nonsense mutation (115C→T) was identified in a four-generation family containing 8 affected males, and results in the conversion of glutamine to stop codon (Q39X).

Interestingly, a total of 9 mutations in 9 independent families with NHS have been reported to date, all of which are truncating and are predicted to result in an aberrant protein product (table 4.5). Families who have truncating mutations in exon 1a, 2 or 3 of the NHS gene (44 %; Burdon *et al.*, 2003; this study; Ramprasad *et al.*, 2005), are predicted to have a loss of the major NHS isoform (isoform A, figure 4.9). These families may still express full-length isoform B, however this requires further investigation. Those families who have truncating mutations in exons 4-8, are predicted to express truncated forms of all 5 *NHS* isoforms (Burdon *et al.*, 2003; this study). Isoform A therefore appears to be the major isoform responsible for disease in families with mutations in the *NHS* gene.

It should be noted that for one Nance-Horan syndrome family screened for mutations in *NHS* (Burdon *et al.*, 2003) no mutation was identified (Burdon *et al.*, 2003). The authors conclude that the mutation may reside within either non-coding sequence of the *NHS* gene, or within regulatory sequence. It should not be ruled out however, that NHS may prove to be heterogeneous, or that the family suffer from a highly related X-linked developmental syndrome such as Oculofaciocardiodental syndrome (OMIM 300166).

Although a mutation within the *NHS* gene has not yet been identified in the CXN family, it is still possible that CXN and NHS are allelic. An attempt to identify alternate 5' ends or cryptic exons of the *NHS* gene may facilitate the identification of a mutation

in the CXN family, and in the family with NHS for which a mutation has not yet been found (Burdon *et al.*, 2003). Upstream and other regulatory regions of the gene should also be considered as mutation sites leading to disease.

At present, there does not appear to be any obvious phenotype/genotype correlation. The affected male of NHS family 3 in this study was reported to be severely developmentally delayed and was found to harbour a deleterious mutation in exon 6 of the *NHS* gene. Such a mutation in exon 6 is predicted to affect all known isoforms of the *NHS* gene (A-E, figure 4.9). In line with this observation, both NHS families 2 and 4, which have truncating mutations in exon 1a (predicted to affect isoform A only), there has been no report developmental delay. Contrary to this, NHS family 1 were found to harbour a truncating mutation in exon 6 and were also not reported to exhibit developmental delay. Therefore, questions arise such as: are truncated versions of the NHS protein expressed and if so, does the number of additional novel amino acids arising from a frameshift have varying affects within the cell? Or, are genetic modifiers involved? The latter appears more likely as Burdon and colleagues reported one family with NHS, in which two of the four affected males had severe intellectual impairment (Burdon *et al.*, 2003). Further acquisition of families with NHS and identification of potential *NHS* mutations, may help to highlight any potential genotype/phenotype correlation.

The *NHS* gene currently consists of 11 exons, of which 10 are coding. At present, there are 5 *NHS* isoforms on the database (figure 4.9). Isoform A codes for the full-length protein of 1,630 amino acids (aa). Isoforms B-E code for a protein of 1,335 aa, 1,474 aa, 1,453 aa, and 1,473 aa, respectively. Four conserved putative monopartite nuclear localization signals were detected at amino acid positions 371-379 (RRRKLRRRK), 438-444 (PSRRRIR), 822-825 (RKPK), and 1026-1034 (PGGSKRKPK) of the major isoform (isoform A, Burdon *et al.*, 2003). Such monopartite nuclear localization signals imply that NHS may have a functional role within the cell nucleus.

Expression of *NHS* in this thesis was determined through PCR analysis of a human fetal multiple tissue cDNA panel, and mouse tooth cDNA (P0 and P2). Primers designed to amplify the human *NHS* gene, allowed for possibly amplification of all predicted *NHS* isoforms (isoforms A-E, figure 4.9). *NHS* transcripts were detected in mouse tooth

cDNA at P0 and human fetal brain, thymus, lung, and kidney. Primers used to amplify the mouse tooth cDNA were designed within exon 6 of the NHS gene. Thus, possible contamination of the mouse tooth cDNA with genomic DNA could not be verified. It is plausible therefore, that the PCR product detected at P0 is in fact a result of genomic contamination, and primers spanning an intron of the mouse *Nhs* gene should be designed to resolve this. It is possible that the mouse tooth cDNA at P1 may be of poor quality, explaining the reason for not being able to detect an *Nhs* transcript at P1. However, failure to amplify an *Nhs* transcript from mouse P1 tooth cDNA may be a result of the *Nhs* gene being switched off. A house keeping gene could also be used as a control to determine the quality of the P0 and P1 mouse tooth cDNA samples.

Elegant expression data for the *NHS* gene was reported by Burdon and colleagues (2003). RT-PCR analysis of all five *NHS* transcripts (A-E, figure 4.9), detected in adult and fetal brain, lens, retina, retinal pigment epithilium, placenta, lymphocytes and fibroblasts (Burdon *et al.*, 2003). Transcript levels of *NHS* in retinal pigment epithilium, placenta, lymphocytes and fiboblasts were reported to be very low. The developmental expression pattern of *NHS* in mouse brain was investigated by probing a mouse northern blot containing brain RNA from various developmental stages. A probe was generated from the 3' end of the *Nhs* gene, thereby allowing for detection of all possible *Nhs* isoforms. Their results showed the mouse *Nhs* gene to be broadly expressed throughout brain development and in adult mouse brain from E17.5 until age 12 months, with transcript levels varying. In addition, Burdon and colleagues performed *in situ* hybridisation of mouse embryo sections using probes designed to exon 1 of isoform A and exon 6 (all *Nhs* isoforms). Expression was detected in midbrain, lens, retina, tooth primordia, the olfactory epithelium, the whisker follicles, the choroid plexus, and heart. Expression in the choroid plexus and heart were only detected using the probe designed from exon 6 of the *Nhs* gene. In light of this, and the absence of a mutation within the CXN family in which cardiac anomalies are documented, a mutation in the CXN family may reside in some regulatory region of the isoforms, which do not contain exon 1a-3.

Expression of *Nhs* (driven by the *Nhs* promoter) throughout development, was studied further by Burdon and colleagues using a mouse line containing a *lacZ* reporter-gene cassette inserted in the *Nhs* locus. Insertion of the cassette occurred within intron one

and therefore was likely to be driven by promoters upstream of exon 2. Consequently, any observed expression was most likely representative of isoform A. Strong expression was detected in the developing lens and relatively high expression was also detected in some tissues of the developing and early postnatal brain. Expression in the olfactory bulbs, olfactory epithilium and the limbic system were most prominent and expression in the heart, apical ectoderm ridge and the olfactory tubercle was also observed.

The preliminary expression studies of *NHS* in this thesis, did not detect the presence of any of the five *NHS* transcripts (A-E, figure 4.9) in human fetal heart. Northern blot results reported by Burdon and colleagues in multiple human fetal and adult tissues are not shown and tissues in which low levels of expression were detected are not specified. Interestingly however, expression of *Nhs* (all potential isoforms) in heart was detected through mouse *in situ* hybridisation and analysis of a mouse *Nhs lacZ* insertion mutant. This may imply that the human fetal heart cDNA sample used in this study was of poor quality. Since cardiac anomalies have been reported in four of the six affected males of the CXN family and CXN and NHS may still prove to be allelic, it is important to determine whether or not the *NHS* gene is expressed (and which isoforms) in human fetal and adult heart. The degree of conservation between vertebrate species (in particular mouse, rat and human, see chapter 5) and the results of *Nhs* expression in the mouse, imply that the human *NHS* gene is likely to be expressed in human heart and therefore remains a strong candidate for CXN.

Prior to cloning the NHS gene, Burdon and colleagues were able to refine the disease interval for NHS to ~1.3 Mb flanked by markers DXS1195 and DXS999. Toutain and colleagues also reported the same refinement of the NHS disease interval in a different family with NHS (Toutain *et al.*, 2002). Interestingly, the truncating mutations identified within exon 1a of families with NHS reside outside the refined NHS disease interval, emphasising the need for caution when refining disease intervals using multiple pedigrees for positional cloning.

- Chapter Five -

The *NHS* gene family, orthologs and paralogs

5.1 Introduction

The *NHS* gene, which was found to harbour truncating mutations in all NHS families in this study, codes for a relatively large full-length protein of 1,630 amino acids (isoform A, figure 4.9). To date, no sequence similarity to any known protein or class of protein has been reported for the NHS protein. Thus, there are no clues as to the function of the NHS protein. Only four putative monopartite nuclear localisation motifs have been described (Burdon *et al.*, 2003). To help gain a greater insight into the function of the NHS protein, NHS orthologs were sought. In particular, a mouse NHS ortholog was identified as part of the positional cloning strategy (chapter four) since the Xcat mouse was believed to be a model for Nance-Horan syndrome (section 3.1.3).

5.2 Results

5.2.1 NHS orthologs

Bioinformatic analysis of the NHS protein has revealed NHS orthologs in a number of vertebrate species (for methods see section 2.7). Using the NHS partial and full-length protein sequence (isoform A) as query, BLAST algorithms were performed against the protein database at NCBI.

NHS orthologs were found in mouse (Mus *musculus*), Rat (Rattus *norvegicus*), Chimpanzee (Pan *Troglodytes*), Dog (Canis *familaris*), Orangutan (Pongo *pygmaeus*, partial), chicken (Gallus *gallus*, partial), Cow (Bos *Taurus*, partial), frog (Xenopus *laevis*, partial), zebrafish (Danio *rerio*, partial), and pufferfish (Tetradon *nigroviridis*, partial). A full alignment of the NHS protein in all these species and humans (Homo *sapiens*) can be seen in figure 5.1, which highlights the conserved domains of the NHS protein. Orthologs in yeast and plants were sought, but none have been identified to date. The level of protein identity between the human NHS protein and its orthologs is summarised in table 5.1.

Ortholog species	**Protein length (aa)**	**Protein identity to Homo *sapiens* NHS protein (%)**	**Protein similarity to Homo *sapiens* NHS protein (%)**	**Genomic contigs and/or cDNAs accession no.**
Canis *familaris**	1628	91	94	NW_879562
Mus *musculus**	1626	88	91	CN531035/AI874417
Rattus *norvegicus**	1627	87	91	NW_048040
Pan *Troglodytes**	1606	95	95	NW_121679
Pongo *pygmaeus*	856 (partial)	99	99	CAH89962
Gallus *gallus*	1328 (partial)	71	80	NW_060228
Bos *Taurus*	1349 (partial)	90	93	BF776631
Xenopus *laevis*	1443 (partial)	62	74	AAH85022
Danio *rerio*	1256 (partial)	55	67	AL926872
Tetradon *nigroviridis*	1527 (partial)	39	52	CAF95442

Table 5.1 Orthologs of the NHS protein and the percentage sequence identity with human NHS protein. * denotes identified full length orthologs for which genomic structures are deduced in fig 5.2.

The NHS protein is highly conserved at the very N- and C-terminus across species (figure 5.1). In the middle of the protein, the level of conservation remains relatively high, with two blocks of highly conserved sequence across all species residing at amino acid positions 550-653 and 924-1043, of the human protein sequence (figure 5.1). A few additional, smaller blocks of highly conserved sequence across all species can be seen throughout the length of the protein sequence (figure 5.1).


H. 1 MPFAKRIVEPQWLCRQRRPAPGPAVDASGGSAEPPPPLQPPGRRDLDEVEAPGPEEPARAVPAPSGLPPPPPP-LPAPADQTQPPHGEASVAGEESTAGIPEAAPAAGEA 109
P. 1 MPFAKRIVEPQWLCRQRRPAPGPAVDASGGSAEPPPPLQPPGRRDLDEVEAPGPEEPARAVPAPSGLPPPPPP-LPAPADQTQPPHGEASVAGEESTAGIPEAAPAAGEA 109
C. 1 MPFAKRIVEPQWLCRQRRPAPGPAEDANGGSADPPPPLQPPGRRD--EAVAPGPEDPPRAPPAPPG-PPPPPP-LPAPTDQAQPPHGEAPAAGEESAAGVAEAASAAGEA 107
M. 1 MPFAKRIVEPQWLCRQRRPAPGPDEDTSGGSVEPPPPLQPPGRRE--EAEAPEPEEPPRVPPAPLPLPPLPPS-LPAPAEQDQQPPSEAREAGEESVAGVSEAASAPGET 107
R. 1 MPFAKRIVEPQWLCRQRRPAPGPDEDTSGGSVEPPPPLQPPGRRE--EAEATEPDEPPRVPPAPLPLPPPPPPSMPAPAEQDQQPPSEAREAGEESVAGVPEAASATGEA 108
B.
Pp.
G.
T.
X.
D.
Exon 1
Exon 2
H. 110 SSAAAAAAVLLMLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHRRVWALQGKLGGVQRVLSTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPATR 219
P. 110 SSAAAAAAVLLMLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHRRVWALQGKLGGVQRVLSTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPATR 219
C. 108 SSAAAAA-VLLMLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHRRVWALQGKLGGVQRVLGTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPATR 217
M. 108 SEASAEA-VLLMLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHRRVWALQGKLGSVQRVLSTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPATR 217
R. 109 SEAAAEA-VLLMLDLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHHRVWALQGKLGSVQRVLNTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPVTR 218
B.
Pp.
G.
T. VSNLDIESKLSLHYQAPWHQQHNVFHPCSR
X. VSNLDAESKLDVFYKSSWHQQRNIFLPITR
D. VSNLDIESKLTSHYQAPWHQQHNLFHTCTR
Exon 2
Exon 3
H. 220 PPCVEELHRHARQSLQALRR----------EHRSRSDRREQRAAAPLSIA-APPLPAYPPAHSQRRREFKDRH---FLT---------- 284
P. 220 PPCVEELHRHARQSLQALRRGDRSWAWGLLRETTRRLSHPEHRSRSDRREQRAAAPLSIA-APPLPAYPPAHSQRRREFKDRH---FLT---------- 304
C. 218 PPCVEELHRHARQSLQALRR----------EHRSRSDRREQRAAAPLSVA-APPLPAYPPAHSQRRREVKDRH---FLT---------- 282
M. 218 PPCVEELHRHARQSLQALRR----------EHRSRSDRREQRAAAPLPIA-APPLPAYPPAHSQRRREAKDRH---FLT---------- 282
R. 219 PPCVEELHRHARQSLQALRR----------EHRSRSDRREQRAAAPVPIA-APPLPAYPPAHSQRRREAKDRH---LLT---------- 283
B.
Pp.
G.
T. PPCVEELHRTAQFSLRALHRD----------EPAQRSASRERNRVTIAISV-APPMPTFPSPHSIRRQQRSRLA---RAQ----------
X. PICVEELHRHAKQSLRSMRK----------EQRIRADNRERRFQGSI-FT-APPLPTYPAVNNLKRQEIKERH---VVQFNRTRSHSPTECCHLTPWSRK
D. PACLEELHRHAKLNLRALHR----------DHLSRSTSRE-RSRVTISISVAPPMPTFPSTHKLRRREQRGRHSRMYQR----------


Exon 4
Exon 4
Exon 5
Exon 5
Exon 6


H. 550 ------SPQHKLSERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPH 653
P. 526 ------SPQHKLSERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPH 629
C. 548 ------SPQHKLNERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPH 651
M. 548 ------SPQQKLSERGRSRLSRMAADSGSCDISSNSDTFGSPVHCISTASVLLSSHMDQKEDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPN 651
R. 549 ------SPQQKLSERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTASVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPN 652
B. ------SPQHKLNERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPH
Pp. ------SPQHKLSERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPH
G. ------SPEHSIGERGRSRLSRMA-DSGSCEISSNSDTFGSPIHSISTAGVLLSSHMDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPN
T. GCGGGGSPEHSWMERGRSRLPRKA-DMGSCEISSSSDTFSSPIHSVSAAGVL-GSHVDHKEDHQSSSGNWSGSSSTCPSQTSESIPPPSSPPLTGSSHCDSELSLNTVPN
X. ------SPQLHLNERGRSRLSRM-ADSGSCDISSNSDTFGSPVHSISAAGVLLSTHIDQKDDHQSSSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPN
D. ------SPDHTWIERGRSRLPRKV-DMGSCEISSSSDTFSSPIHSASTTGVL-GSQIDHKEDHQSSSGNWSGSSSTCPSQTSETLPPAASPPLTGSSHCDSELSLNTGSH
H. 654 ANEDASVFVTEQY-NDHLDKVRGH--RANSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSLG--CPSFTSMATYD-----SFLEKSPSDKAD 751
P. 630 ANEDASVFVTEQY-NDHLDKVRGH--RANSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSRG--CPSFTSMATYD-----SFLEKSPSDKAD 727
C. 652 ANEDANVFVTEQY-NDHLDKVRGH--RANSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSLG--CPSFTSMATYD-----SFLEKSPSDKAD 759
M. 652 ANEDASVFVTEQY-NDHLDKVRGH--RTNSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSLG--CPSFTSMGTYD-----SFLEKSPSDKAD 759
R. 653 ANEDASVFVTEQY-NDHLDKVRGH--RTNSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPEHPKADSLG--CPSFTSMGTYD-----SFLEKSPSDKAD 760
B. ANEDTSVFVTEQY-NDQLEKVRGH--RTNSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSLG--CPSFTSMATYD-----SFLEKSPSDKAD
Pp. ANEDASVFVTEQY-NDHLDKVRGH--RANSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERPKADSLG--CPSFTSMATYD-----SFLEKSPSDKAD
G. ANEDSSVFITEQY-GDHVDKVRGH--RASSFTSTVADLLDDPNNSNTSDSEWNYL--HHHHDASCRQDFSPERTKADSLG--CPSFTSMGTYD-----SFLEKSPSDKAD
T. ATDDGFSLDPSYH-SDLRPQGQGH--RSSSFTSSATDQLDDAGVSTASEGEWTYPPDQDQTDQDTETDQSRTLNRCQDLP---REYNSKEGQENQVCFSDIKSSNVDK-D
X. AGEDSNVFITEQYSDDHIDSIRGH--RASSFTSTVADLLDDP-NSNTSDSEWNYL--HHHHDASCHQDFSPQHLKTDRLA--CPSFTSMGTYD-----SFLEKPASEKAD
D. VIDDNTGFIIDPY---HIDKPQGQGQRSNSFTSSATDQMDDAGVSTASDGEWNCA--QDQQDQSCTQDFSPKRSREDESGVSCPSYSSGETYE-----SFSDQASARKSE
H. 752 TSSHFSVDTEGYYT-SMHFDCGLKGNKSYVCHYAALGPENGQGVGASPGLPDCAWQDYLDHKRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKEPKIS 855
P. 723 TSSHFSVDTEGYYT-SMHFDCGLKGNKSYVCHYAALGPENGQGVGASPGLPDCAWQDYLDHKRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKEPKIS 826
C. 760 TSSHFSVDTEGYYT-SMHFDCGLKGNKSYVCHYAALGPENSQGIGVPPTLPDCTWQDYLDHRRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKEPKIS 863
M. 760 TSSHFSVDTEGYYT-SMHFDCGLKGNKTYVCHYAGLGPENGQGVGVPPALPDCAWQEYLDHRRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKEPKIS 863
R. 761 TSSHFSVDTEGYYT-SMHFDCGLKGNKTYVCHYAGLGPENGQGVGVPPALPDCAWQEYLDHRRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKEPKIS 864
B. TSSHFSVDTEGYYT-SMHFDCGLKGSKNYVCHYAALGPENSQGIGVPPALSDCAWQDYLDHRRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADT-----SEKKEPKTS
Pp. TSSHFSVDTEGYYT-SMHFDCGLKGNKSYVCHYAALGPENGQGVGASPGLPDCAWQDYLDHKRQGRPSISFRKPKAKPTPPKRSSSLRKSDGNADI-----SEKKESKIS
G. TSSHFSVDTEGYYT-SMHFDCGLKGNKSYICNYAAGSESGQMASMTSSLADCTWQECISHRRQGRQGISLKKPKAKPAPPKRSSSLRKSEGSMDL-----PEKKEPKIS
T. PSSNYSSDTEGFYTSSVHFGECNQGYRGYMYNYADPGPDCAQSNTTAAPYPQHMADSRGGTMTLGRTCRSLRKPKVKPPPPKRTSSLKEPVCSVDVETNTQTEQDQPKMA
X. TGSHFSVDTEGYYT-SMHFDCGLNTGRSYIYNYAPTGSERDQGRGISSMNTDNAWQCYMDQENEGNQNISFKKPKARPTPPKRSSSLRKSDNNTDI-----LEKKEPKIV
D. M--HYIVDTEGFYS-SMHFDSGLKGSRSYY-NYAAIDPDCGPSGNIAPGMTEYPQPEYRATRTLGRPCLSLRKPKAKPPPPKRSSSLKETSSDVNV-----PEQNEPKIT

```
H.   856 SG-QHLPHSSREMKLPL--DFANTPSRMENANLPTKQEPSW---INQSEQGIKEPQLDASDIPPFKDEVAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI 957
P.   827 SG-QHLPHSSREMKLPL--DFANTPSRMENANLPTKQEPSW---INQSEHGIKEPQLDASDIPPFKDEVAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI 928
C.   864 SG-QLLPHSSREMKLPL--DFSNTPSLMENANLAIKQESSW---MNQSEHGIKEPQLDTSDIPPFKDEGAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI 965
M.   864 SG-QHLPHSSREMKLPL--DFSNTPSRVENANLPAKLDSSW---INQSEHAIKEPQLDTPDISPFKDEGAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI 965
R.   865 SG-QHLPHSSREMKLPL--DFSNTPSRVENANLPAKLDSSW---ISQSEHAIKEPQLDTTDISPFKDEGAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI 966
B.       SG-QLLPHSSREMKLPL--DFSNTPSLMENANLPTKPEPSW---MNQNEHGIKEPQLDTPDVPPFKDEGAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI
Pp.      SG-QHLPHSSREMKLPL--DFANTPSRMENANLPTKQEPSW---INQSEHGIKEPQLDASDIPPFKDEVAESTHYADLWLLN-DLKTND-PYRSLSNSSTATGTTVIECI
G.       SG-QHTSHTAREMKLPL--EFSNTPSRAEASSLPPKPELPW---VSQSDGGMKDTQFDMADMPSFKDDGAEQPHYADLWLLN-DLKSGD-PYRSLSNSSTATGTTVIECI
T.       TE-QEDTLSSTDTNLELEQDLESAPEPLQTACLVAEPLGTWGMGLHETVDIVEPMSFSSADTHSFKDEGAVQSDYADLWLHNTDLKCNNGEYTSMSNSSTATGTTVMECM
X.       CGQQQVIHSSREMKLPL--EMSNIPSRVENFATPNQQEIPW---VNQDDNDLKSTQFLTTDVPSFKYESGEQSHYADLWLVN-DLKTSD-PYRSLSNSSTATGTTVIECI
D.       CE-LPLPLSSREMKLQL--DLADSAGHLETAGLESLGAWGV---ENVRDMLDCSSPFSSSDTHSFKDEGAVQADYADLWLHN-DLKSND-PYRSLSNSSSATGTTVIECI

H.   958 KSPESSESQT--------SQSESRATTPSLPSVDNEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL---- 1047
P.   929 KSPESSESQT--------SQSESRATTPSLPSVDNEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL---- 1018
C.   966 KSPESSESQT--------SQSESRATTPSLPSVDNEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL---- 1055
M.   966 KSPESSESQT--------SQSESRATTPSLPSVDNEYKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL---- 1055
R.   967 KSPESSESQT--------SQSESRATTPSLPSVDNEYKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPQSPGGSKRKPKVPERKSSL-----QQPSL---- 1056
B.       KSPESSESQT--------SQSESRATTPSLPSVDNEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL----
Pp.      KSPESSESQT--------SQSESRATTPSLPSVDNEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPSL----
G.       KSPESSESQT--------SQSGSRATTPSLPSVDSEFKLASPEKLAGLASPSSGYSSQSETPTSSFPT---AFFSGPLSPGGSKRKPKVPERKSSL-----QQPLT----
T.       KSPESSSSSTEASTQALSKASDTRAASPPLPSGD--FKLGSPEKLAGLASPSSGYSSQSETPTSTLPSSSGAFFPGPQSPSTGKRKPKVPERKSSL-------PSLQHFS
X.       KSPESSESQM--------SQSESRATSPSLPSVENEFKLSSPEKLAGLASPSSGYSSQSETPTSSFPT---PFFSGPLSPGGSKRKPKVPERKSSL-----QHPSA----
D.       KSPERSETHT--------CQPRSRPSSPTLPPPESEFKLASPEKLAGLASPSSGYSSQSETPTSSFPS---AFFPGPLSPTSGKRKPKVPERKSSLCSLQQQQLSV----

H.  1048 KDGTISLS---KDLELPIIPPTHLDLSALH-----NVLNKPFHHRHPLHVFTHNKQNTVG-------ETLRSNPPPS------------LAITPTILKSVNLRSINK--- 1127
P.  1019 KDGTISLS---KDLELPIIPPTHLDLSALH-----NVLNKPFHHRHPLHVFTHNKQNTVG-------ETLRSNPPPS------------LAITPTILKSVNLRSINK--- 1098
C.  1056 KDGALSLS---KDLELPIIPPTHLDLSALH-----SVLNKPFHHRHPLHVFTHSKQNPVG-------ETLRSNPPPS------------LAITPTVLKSVNLRSISK--- 1135
M.  1056 KDGALSLS---KDFELPIIPPSHLDLSALH------VLNKPFHHRHPLHVFTH-KQNTVG-------DMPRSNPAPS------------LAITPTVLKSVNLRSISK--- 1133
R.  1057 KDGALSLS---KDFELPIIPPSHLDLSALH------VLNKPFHHRHPLHVFTH-KQNTVG-------DMLRSNPPPS------------LAITPTVLKSVNLRSISK--- 1134
B.       KDGALSLS---KDLELPIIPPTHLDLSALH-----NVLNKPFHHRHPLHVFSHNKQNTVG-------ETLRSNPPPS------------LAITPTVLKSVNLRSISK---
Pp.      KDGTISLG---KDLELPIIPPTHLDLSALH-----NVLNKPFHHRHPLHVFTHNKQNTVG-------ETLRSNPPPS------------LAITPTILKSVNLRSINK---
G.       KDGTMSAS---KDLELPIIPPTHLDLSALH-----SVLNKPFAHRHQLHAFNPSKHSAVG-------EALSPSHPSA------------LAITPSVLKSVHLRAINR---
T.       RDGTSISSGYRRDPDFP-PPPSQLDLNILHGGYIRHTLSHRTHHMHTLHHNKHRATNVVATGAKFVTEVSNANQSASSNLSQTSASTSVSVITPSVPRSVQFHPVSQPTA
X.       KCCNVLGG---KELELPTIPPTNLDLSALH------VLHKPLPYRPHIHAFSHNNQNMSE-------DQFDSNSSTA------------LAITPSVLKSVQLRSVNK---
D.       RDPGISCR---RETDFYAIPPSHLDLSALH-----TRTRQQKVPRCAKDIVTRETKTRFE-------STIAARP-----------------------------------
```

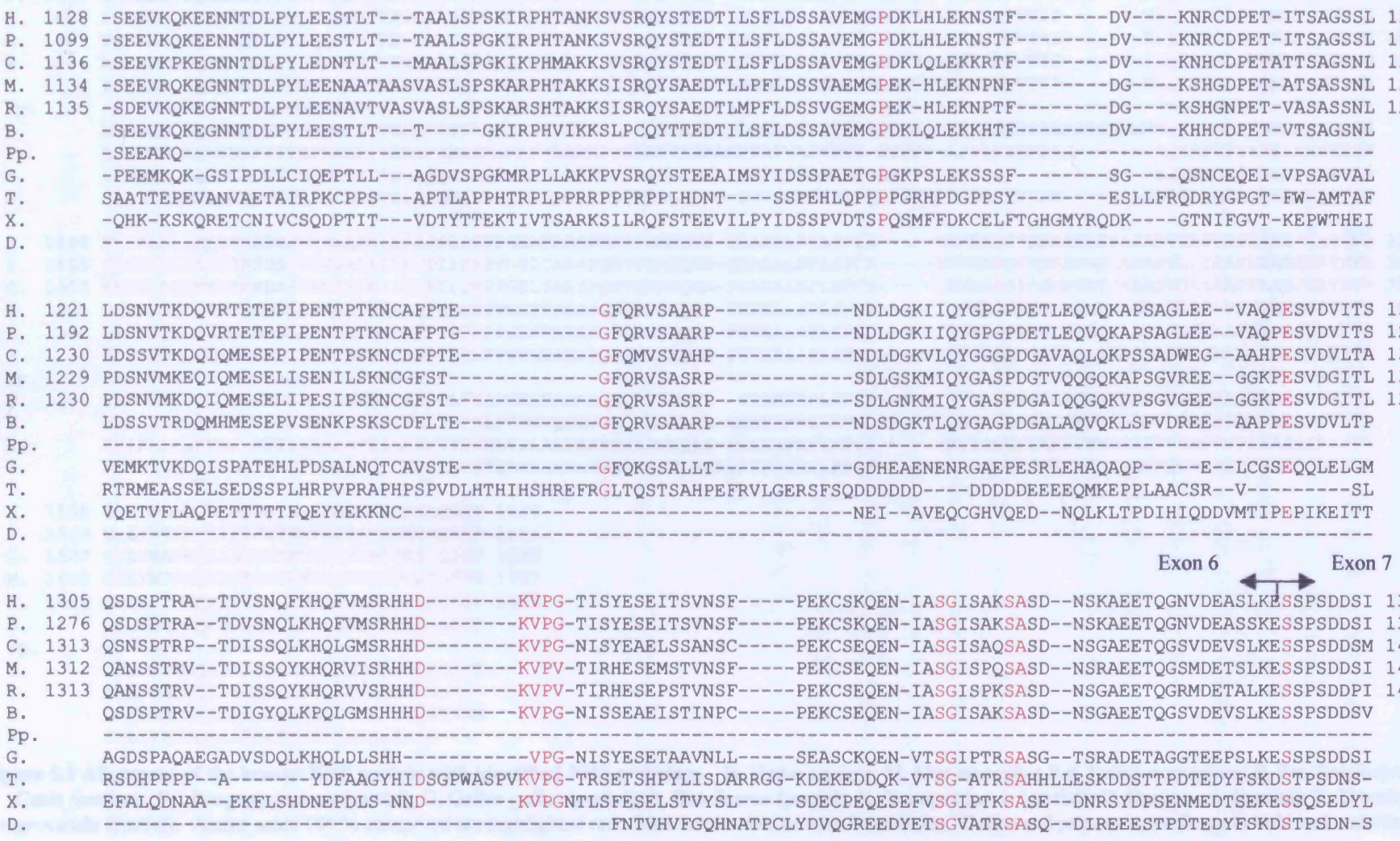

H. 1128 -SEEVKQKEENNTDLPYLEESTLT---TAALSPSKIRPHTANKSVSRQYSTEDTILSFLDSSAVEMGPDKLHLEKNSTF--------DV----KNRCDPET-ITSAGSSL 1220
P. 1099 -SEEVKQKEENNTDLPYLEESTLT---TAALSPGKIRPHTANKSVSRQYSTEDTILSFLDSSAVEMGPDKLHLEKNSTF--------DV----KNRCDPET-ITSAGSSL 1191
C. 1136 -SEEVKPKEGNNTDLPYLEDNTLT---MAALSPGKIKPHMAKKSVSRQYSTEDTILSFLDSSAVEMGPDKLQLEKKRTF--------DV----KNHCDPETATTSAGSNL 1229
M. 1134 -SEEVRQKEGNNTDLPYLEENAATAASVASLSPSKARPHTAKKSISRQYSAEDTLLPFLDSSVAEMGPEK-HLEKNPNF--------DG----KSHGDPET-ATSASSNL 1228
R. 1135 -SDEVKQKEGNNTDLPYLEENAVTVASVASLSPSKARSHTAKKSISRQYSAEDTLMPFLDSSVGEMGPEK-HLEKNPTF--------DG----KSHGNPET-VASASSNL 1229
B. -SEEVKQKEGNNTDLPYLEESTLT---T-----GKIRPHVIKKSLPCQYTTEDTILSFLDSSAVEMGPDKLQLEKKHTF--------DV----KHHCDPET-VTSAGSNL
Pp. -SEEAKQ----------
G. -PEEMKQK-GSIPDLLCIQEPTLL---AGDVSPGKMRPLLAKKPVSRQYSTEEAIMSYIDSSPAETGPGKPSLEKSSSF--------SG----QSNCEQEI-VPSAGVAL
T. SAATTEPEVANVAETAIRPKCPPS---APTLAPPHTRPLPPRRPPPRPPIHDNT----SSPEHLQPPPPGRHPDGPPSY--------ESLLFRQDRYGPGT-FW-AMTAF
X. -QHK-KSKQRETCNIVCSQDPTIT---VDTYTTEKTIVTSARKSILRQFSTEEVILPYIDSSPVDTSPQSMFFDKCELFTGHGMYRQDK----GTNIFGVT-KEPWTHEI
D. ----------
H. 1221 LDSNVTKDQVRTETEPIPENTPTKNCAFPTE----------GFQRVSAARP----------NDLDGKIIQYGPGPDETLEQVQKAPSAGLEE--VAQPESVDVITS 1304
P. 1192 LDSNVTKDQVRTETEPIPENTPTKNCAFPTG----------GFQRVSAARP----------NDLDGKIIQYGPGPDETLEQVQKAPSAGLEE--VAQPESVDVITS 1275
C. 1230 LDSSVTKDQIQMESEPIPENTPSKNCDFPTE----------GFQMVSVAHP----------NDLDGKVLQYGGGPDGAVAQLQKPSSADWEG--AAHPESVDVLTA 1312
M. 1229 PDSNVMKEQIQMESELISENILSKNCGFST----------GFQRVSASRP----------SDLGSKMIQYGASPDGTVQQGQKAPSGVREE--GGKPESVDGITL 1311
R. 1230 PDSNVMKDQIQMESELIPESIPSKNCGFST----------GFQRVSASRP----------SDLGNKMIQYGASPDGAIQQGQKVPSGVGEE--GGKPESVDGITL 1312
B. LDSSVTRDQMHMESEPVSENKPSKSCDFLTE----------GFQRVSAARP----------NDSDGKTLQYGAGPDGALAQVQKLSFVDREE--AAPPESVDVLTP
Pp. ----------
G. VEMKTVKDQISPATEHLPDSALNQTCAVSTE----------GFQKGSALLT----------GDHEAENENRGAEPESRLEHAQAQP-----E--LCGSEQQLELGM
T. RTRMEASSELSEDSSPLHRPVPRAPHPSPVDLHTHIHSHREFRGLTQSTSAHPEFRVLGERSFSQDDDDDD----DDDDDEEEEQMKEPPLAACSR--V---------SL
X. VQETVFLAQPETTTTTFQEYYEKKNC----------NEI--AVEQCGHVQED--NQLKLTPDIHIQDDVMTIPEPIKEITT
D. ----------
Exon 6
Exon 7
H. 1305 QSDSPTRA--TDVSNQFKHQFVMSRHHD--------KVPG-TISYESEITSVNSF-----PEKCSKQEN-IASGISAKSASD--NSKAEETQGNVDEASLKESSPSDDSI 1395
P. 1276 QSDSPTRA--TDVSNQLKHQFVMSRHHD--------KVPG-TISYESEITSVNSF-----PEKCSKQEN-IASGISAKSASD--NSKAEETQGNVDEASSKESSPSDDSI 1366
C. 1313 QSNSPTRP--TDISSQLKHQLGMSRHHD--------KVPG-NISYEAELSSANSC-----PEKCSEQEN-IASGISAQSASD--NSGAEETQGSVDEVSLKESSPSDDSM 1403
M. 1312 QANSSTRV--TDISSQYKHQRVISRHHD--------KVPV-TIRHESEMSTVNSF-----PEKCSEQEN-IASGISPQSASD--NSRAEETQGSMDETSLKESSPSDDSI 1402
R. 1313 QANSSTRV--TDISSQYKHQRVVSRHHD--------KVPV-TIRHESEPSTVNSF-----PEKCSEQEN-IASGISPKSASD--NSGAEETQGRMDETALKESSPSDDPI 1403
B. QSDSPTRV--TDIGYQLKPQLGMSHHHD--------KVPG-NISSEAEISTINPC-----PEKCSEQEN-IASGISAKSASD--NSGAEETQGSVDEVSLKESSPSDDSV
Pp. ----------
G. AGDSPAQAEGADVSDQLKHQLDLSHH----------VPG-NISYESETAAVNLL-----SEASCKQEN-VTSGIPTRSASG--TSRADETAGGTEEPSLKESSPSDDSI
T. RSDHPPPP--A-------YDFAAGYHIDSGPWASPVKVPG-TRSETSHPYLISDARRGG-KDEKEDDQK-VTSG-ATRSAHHLLESKDDSTTPDTEDYFSKDSTPSDNS-
X. EFALQDNAA--EKFLSHDNEPDLS-NND--------KVPGNTLSFESELSTVYSL-----SDECPEQESEFVSGIPTKSASE--DNRSYDPSENMEDTSEKESSQSEDYL
D. ----------FNTVHVFGQHNATPCLYDVQGREEDYETSGVATRSASQ--DIREEESTPDTEDYFSKDSTPSDN-S

Figure 5.1 Alignment of the human NHS protein with identified NHS orthologs. H, Homo *sapiens*; M, Mus *musculus*; Rat, Rattus *norvegicus*; P, Pan *Troglodytes*; C, Canis *familaris*; Pp, Pongo *pygmaeus* (partial); G, Gallus *gallus* (partial); B, Bos *Taurus* (partial); X, Xenopus *laevis* (partial); D, Danio *rerio* (partial); T, Tetradon *nigroviridis* (partial). Amino acids 100 % conserved are highlighted red. Blue arrows denote exon boundaries. Conserved sequence motifs are in bold and underlined.

Stretches of amino acid sequence (490-549, 856-924, and 1121-1395 of the human NHS protein sequence) were found not to be very conserved across all species (figure 5.1). However, the NHS protein sequence of human (Homo *sapiens*), mouse (Mus *musculus*), Rat (Rattus *norvegicus*), Chimpanzee (Pan *Troglodytes*), and Dog (Canis *familaris*), is much more highly conserved throughout the entire length of the NHS protein. Only on addition of the remaining species, in particular the partial protein sequence of frog (Xenopus *laevis*), zebrafish (Danio *rerio*) and the pufferfish (Tetradon *nigroviridis*) is the level of conservation reduced. The pile up of the NHS protein sequences from all species however, highlights important functional domains of the NHS protein sequence.

To date, all the above NHS protein sequences from each species is based solely on prediction. There is little expression evidence currently available to support these predictions, as highlighted in figure 5.2. Full-length protein sequence was identified only for Mus *musculus*, Rattus *norvegicus*, Pan *Troglodytes*, and Canis *familaris* based on genome assemblies. As a result, the genomic structure of each of these predicted genes was examined. Figure 5.2 depicts the genomic structure for Mus *musculus*, Rattus *norvegicus*, Pan *Troglodytes*, and Canis *familaris*.

Limited expression data was available for the Mus *musculus* homologue. Two unigene clusters (Mm.328515 and Mm.336313) were identified for the *Nhs* gene from Mus *musculus*. Unigene cluster 328515 contains two ESTs from adult male kidney and mouse E14.5 retina, which align with exon 1 of the *Nhs* gene (figure 5.2). Additional ESTs were identified in unigene cluster Mm.336313 and align with the 3' end of the *Nhs* gene. A kidney EST AI874417 from unigene cluster Mm.336313 joins exon 4 of the Mus *musculus Nhs* gene with an exon 3b equivalent of the Homo *sapiens NHS* gene. Therefore exon 3b of the Homo sapien *NHS* gene (isoforms C and E; chapter 4 section 4.2.2, figure 4.9) was found to be present in the Mus *musculus Nhs* gene. An additional two EST hits are present in unigene cluster Mm.336313. The first, CB722921 from whole brain, aligns with exon 6 of the *Nhs* gene (figure 5.2). The second EST joins exons 6, 7, and 8 of the Mus *musculus Nhs* gene (figure 5.2), similar to mRNA AK026164 of the Homo sapien *NHS* gene (chapter 4 section 4.2.1, figure 4.1). No EST or mRNA hits were identified for exons 2, 3, or 5 of the

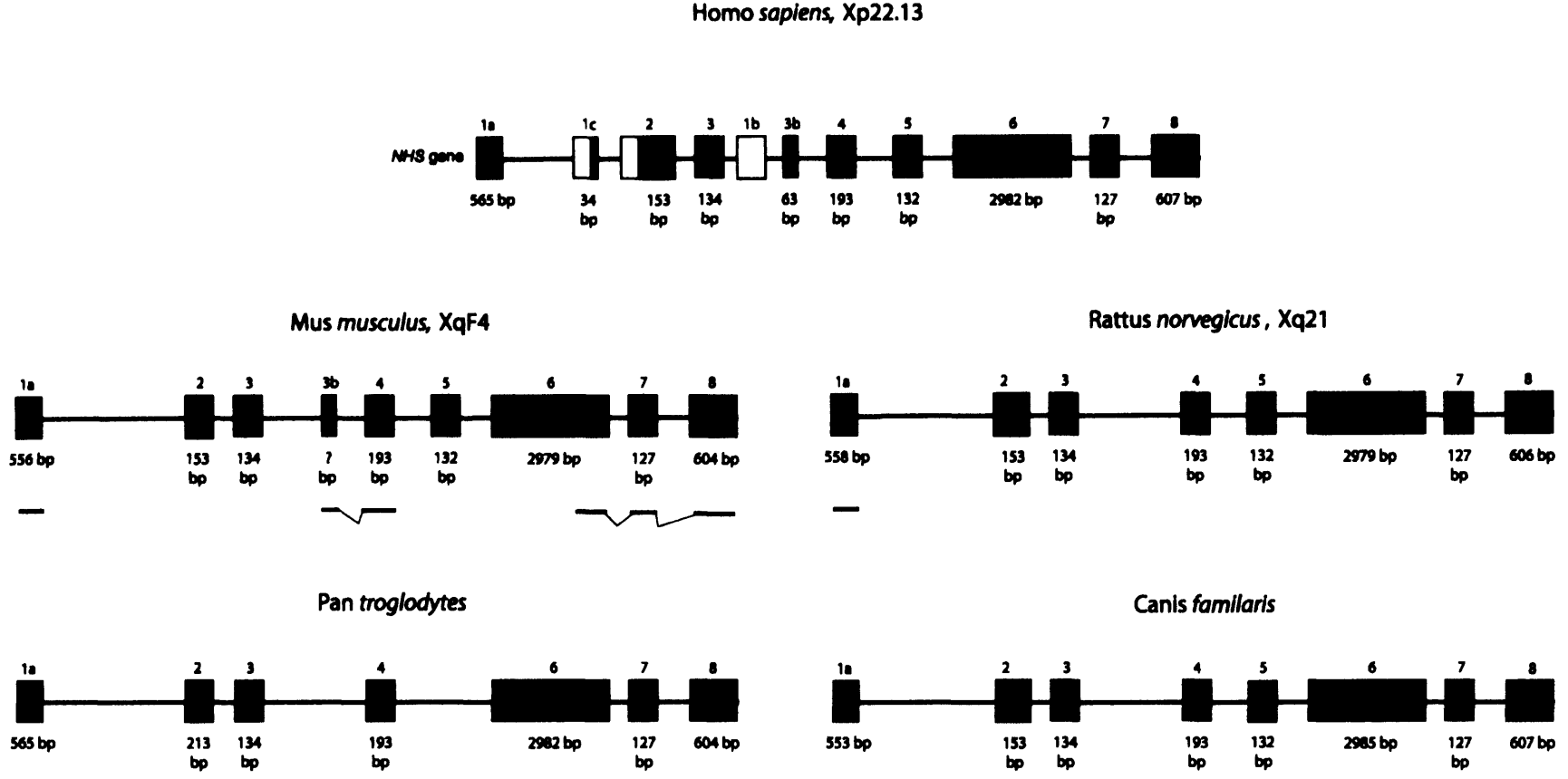


Figure 5.2 Predicted genomic structures of the Nhs gene in Mus *musculus*, Rattus novegicus, Pan *troglodytes*, and Canis *familaris*.

The genomic structure of the NHS gene in Homo *sapiens* is given at the top of the figure for comparison of exons and their sizes. Exons are numbered with respect to the human sequence. EST/cDNA hits are depicted below the corresponding genomic structures (for data on human NHS expression and isoforms see chapter four). Genomic structures for each of the *NHS* orthologs are not to scale.

Mus *musculus Nhs* gene. Further, the alternative 5' exons observed in isoforms B-E of the Homo *sapiens NHS* gene (chapter 4 section 4.2.2, figure 4.9) have not yet supported by ESTs in the Mus *musculus Nhs* gene.

The deduced genomic structures of the Rattus *norvegicus*, Pan *Troglodytes*, and Canis *familaris Nhs* genes can also be seen in figure 5.2. Exon sizes amongst the various species appears to be highly conserved. Sizes of exons 3, 4, and 5 are identical amongst the Homo *sapiens*, Mus *musculus*, Rattus *norvegicus*, Pan *Troglodytes*, and Canis *familaris NHS* genes. Exon 2 differs only in Pan *Troglodytes*, although this may prove to be an error as the sizes of exons 1 and 3-7 are identical between Homo *sapiens* and Pan *Troglodytes*. Furthermore, exon 2 is 100 % conserved between these two species (amino acids 189-240 of the Homo sapien NHS protein, figure 5.1), except for the additional sequence in Pan *Troglodytes*, which is absent in all other species. Currently, there is no expression data to confirm the presence of these predicted exons in Rattus *norvegicus*, Pan *Troglodytes*, and Canis *familaris*, and the genomic structures (figure 5.2) are based purely on analysis of genome assemblies for each of these species.

5.2.2 NHS paralogs

After identifying orthologs of the human NHS protein through bioinformatic analysis, significant homology to sequence KIAA1357 located on chromosome 6q24.1 (www.ncbi.nlm.nih.gov/, using BLAST algorithms) was detected. This sequence was predicted to be a novel gene by computational analysis of supporting mRNA and EST evidence (including UniGene cluster Hs. 170162) from multiple tissues, including adult and fetal brain, fetal eyes, and adult lens, kidney, liver, and intestine.

Inspection of the genomic interval (http://genome.ucsc.edu) revealed the structure of this novel gene, hereafter named *NHSL1* (*NHS-Like 1*), on chromosome 6, with striking resemblance to the genomic structure of the *NHS* gene (figure 5.3).

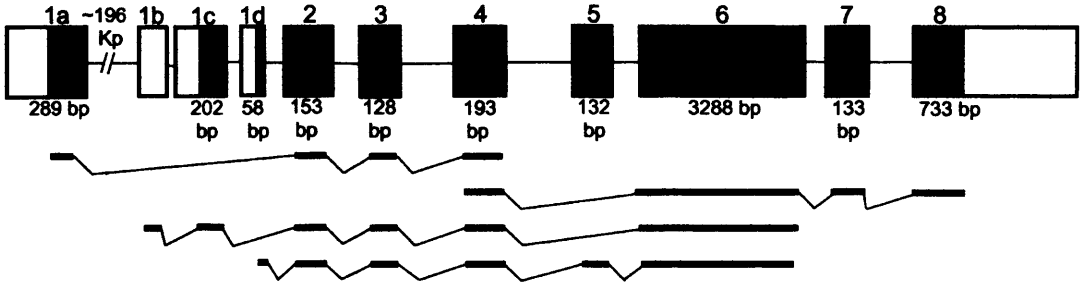


Figure 5.3 *NHSL1* genomic structure.

Blackened squares indicate the open reading frame (ORF), with sizes of coding portions given below. Genomic structure not to scale.

NHSL1 is predicted to be alternatively spliced with at least four alternative 5' exons (1a-1d) and a cryptic exon 5 (figure 5.3). Exon 1a lies 196 Kb upstream of exon 2. Similarly, exon 1 of *NHS* is found 350 kb upstream of exon 2. In addition, exon 1a of *NHSL1* has sequence homology at the protein level to NHS exon 1a (figure 5.5). Exons 2, 4, and 5 of the *NHS* and *NHSL1* genes are identical in size and show a high degree of protein homology, however, the large exon 6 of both genes is divergent. Exon 3, again is similar in size in both genes but encodes different protein sequences.

Interestingly, *REPS1* and *REPS2* lie telomeric to *NHSL1* and *NHS* respectively. REPS1 and REPS2 are close paralogs sharing 32 % protein sequence identity, implying that Xp22.13 and 6q42.1 are derived from a duplicated genomic sequence.

The identification of *NHSL1* prompted a wider search (through bioinformatic analysis), for additional *NHS* paralogs using conserved domains between the protein sequences of NHS and NHSL1 as query. This resulted in significant homology to sequence BC033261 mapping to chromosomal region Xq13.1. Further computational analysis of Xq13.1 revealed a genomic structure highly similar to that of the *NHS* and *NHSL1* genes (figure 5.4). This newly identified paralog on Xq13.1 (named *NHS-Like 2*, *NHSL2*) has supporting mRNA and EST evidence (UniGene cluster Hs. 397836) from brain, mammary gland, and tongue. Exon 1a, like *NHS* and *NHSL1*, is found far upstream of exon 2, approximately 221 Kb. *NHSL2* consists of 8 putative coding exons, predicted to code for a 1189 amino acids and like *NHSL1*, EST evidence suggests splicing out of exon 5 (figure 5.4). In addition, from available expression data, it appears that exons 2-4 are also spliced out of one isoform (figure 5.4). To date, no

alternative 5' exons have been identified for the *NHSL2* gene and there is no supporting expression data for exon 2, although exon 2 of *NHS* and *NHSL1* is highly conserved at the protein sequence level with the *NHSL2* gene. Exon 3, which is divergent between *NHS* and *NHSL1*, is also different in *NHSL2*, although there is no EST or mRNA supporting evidence to confirm its expression. However, the splice sites of *NHSL2* and *NHS* are similar for exon 3 implying that this particular exon is likely to form exon 3 of the *NHSL2* gene.

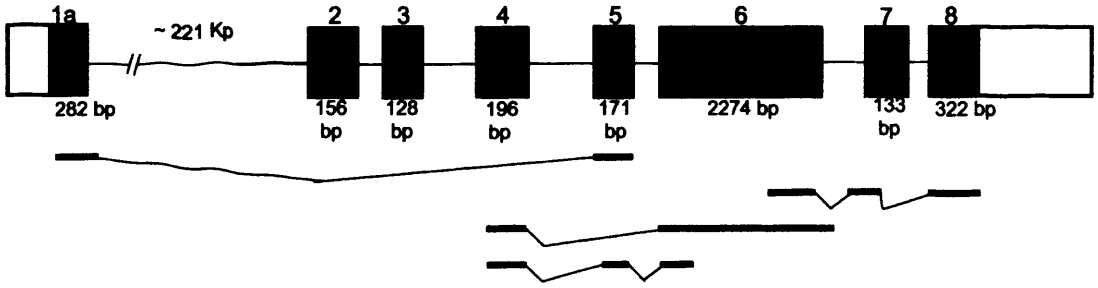


Figure 5.4 *NHSL2* genomic structure.

Blackened squares indicate the open reading frame (ORF), with sizes of coding portions given below. Genomic structure not to scale.

A pile up of the protein sequences for NHS, NHSL1 and NHSL2 (figure 5.5), revealed regions of homology (figure 5.6) indicating specific protein domains for this protein family. In particular, sequence motifs SPSSGYSSQSXTPT, PXVPERKSSL, and RTTEDLFXXIHRSKRKXLG. These motifs were used as query sequence to try and identify any additional family members of the *NHS* gene family. As a result of this additional search, an alignment of the sequence motif SPSSGYSSQSXTPT with a partial human brain mRNA AB040955 mapping to chromosomal region 1p35.1, was identified (figure 5.7). The sequence motif SPSSGYSSQSETPT in NHS is also conserved across species (located within exon 6, figure 5.1).

Homology domain 1

```
NHS    (Exon 1)     1 MPFAKRIVEPQWLCR 15
NHSL1  (Exon 1a)    1 MPFHQRTLEPARLRR 15
NHSL2  (Exon 1a)    1 MPFYRRTVVPQRLC- 14
                        :  :

NHS    (Exon 1)   123 DLCAVSNAALARVLRQLSDVARHACSLFQELESDIQLTHRRVWALQGKLGGVQR 176
NHSL1  (Exon 1)    31 SLEQVSSHALGCLLAQLADLSRCAGDIFGELEGQAAALGHRTAALHRRLDALQA  84
NHSL2  (Exon 1)    24 ELRDVSHLAALSLLRQLADLCGHSLALLEDLEGHLLALGRRTDSLYRRTVRLRR  77
```

Homology domain 2

```
                          :    : :.  :  :: :  ..     : . :   :   ::

NHS    (Exon 2)   190 VSNLDIES-KLSVYYRAPWHQQRNIFLPATRPPCVEELHRHARQSLQALRRE 240
NHSL1  (Exon 2)    98 VSNLDEES-RWTVHYTAPWHQQENVFLPTTRPPCVEDLHRQAKLNLKSVLRE 148
NHSL2  (Exon 2)    95 AANSGRENATATAHSRSSWRQPVNVFLSSGRPPSVEELLREAQLNLQSLLQE 146
```

Homology domain 3

```
       .:  .  .   :.:  :. :    :  .:    .  :   . :   ::: :

NHS    (Exon 4)   310 STLDKQTNWSKALPLPTPEEKMKQDAQVISSCIIPINV 347
NHSL1  (Exon 4)   216 SDLNTQTNWTKSLPLPTPEEKMRQQAQTVQADVVPINI 253
NHSL2  (Exon 4)   215 TTADKQTAWNSLFPLPILEEKRWPQLCSTQSDIVPINI 252
       :  :.    .. :           :    .: ::   :
```

Homology domain 4

```
NHS    (Exon 5)   348 TGVGFDREASIRCSLVHSQSVLQRRRKLRRRKTISGIPRRVQQEI 392
NHSL1  (Exon 5)   254 TGENFDRQASLRRSLIYTDTLVRRPKKVKRRKTITGVPDNIQKEL 298
NHSL2  (Exon 5)   266 TGQQFDKHASLRHSLFNTETAVNPKSTLRRRRTIIGFSNFSQRDQ 310
             :.  :    . ::: :.   .::  :    ..    ::
```

Homology domain 5

```
NHS    (Exon 6)   818 SISFRKPKAKPTPPKRSSSLRK 839
NHSL1  (Exon 6)   739 NISLKKAKKPPLPPSRTDSLRR 760
NHSL2  (Exon 6)   571 SISLRKAKKKPSPPTRSVSLVK 592
       .  :: .       . :    :
```

Homology domain 6

```
NHS    (Exon 6)   972 SRATTPSLPSVDNEFK-LASPEKLAGLASPSSGYSSQSETPTSS 1014
NHSL1  (Exon 6)   897 SRATMPQVPGGSVKPK-IMSPEKSHRVISPSSGYSSQSNTPTAL  939
NHSL2  (Exon 6)   695 SRATTPSQLSIEVEAREISSPGRPPGLMSPSSGYSSQSETPTPT  738
             .  . . : : :    :   :          :   .
```

Homology domain 7

```
NHS    (Exon 6)  1027 GGSKRKPKVPERKSSL 1042
NHSL1  (Exon 6)   953 GKGKPKPKVPERKSSL  969
NHSL2  (Exon 6)   753 SSVRVRPVVPERKSSL  768
          : :
```

Homology domain 8

```
NHS    (Exon 7-8) 1418 RTTEDLFAVIHRSKRKVLGRKDSGDMS 1444
NHSL1  (Exon 7-8) 1428 RTTEDLFAAIHRSKRKVLGRRDSDDDH 1454
NHSL2  (Exon 7-8) 1102 RTTEDLFTVIHRSKRKLLGWKEPGSGS 1128
                  :.       :   ::...
```

Homology domain 9

```
NHS    (Exon 8)   1493 NEEFKLLLLKKGSRSDSSYRMSATEILKS 1521
NHSL1  (Exon 8)   1493 SDNFKALLLKKGSRSDTSARMSAAEMLKN 1521
NHSL2  (Exon 8)   1137 NDDFKALLQKKGSKATPRSRPSAAELLKT 1165
                     .::          :: .      : :  :
```

Figure 5.5 Alignment of NHS, NHSL1, and NHSL2 protein sequences.

Amino acids 100 % conserved are highlighted in red. Homology domains 1-9 can be seen schematically in figure 5.6. Protein sequences for NHSL1 and NHSL2 are shown in appendix A.

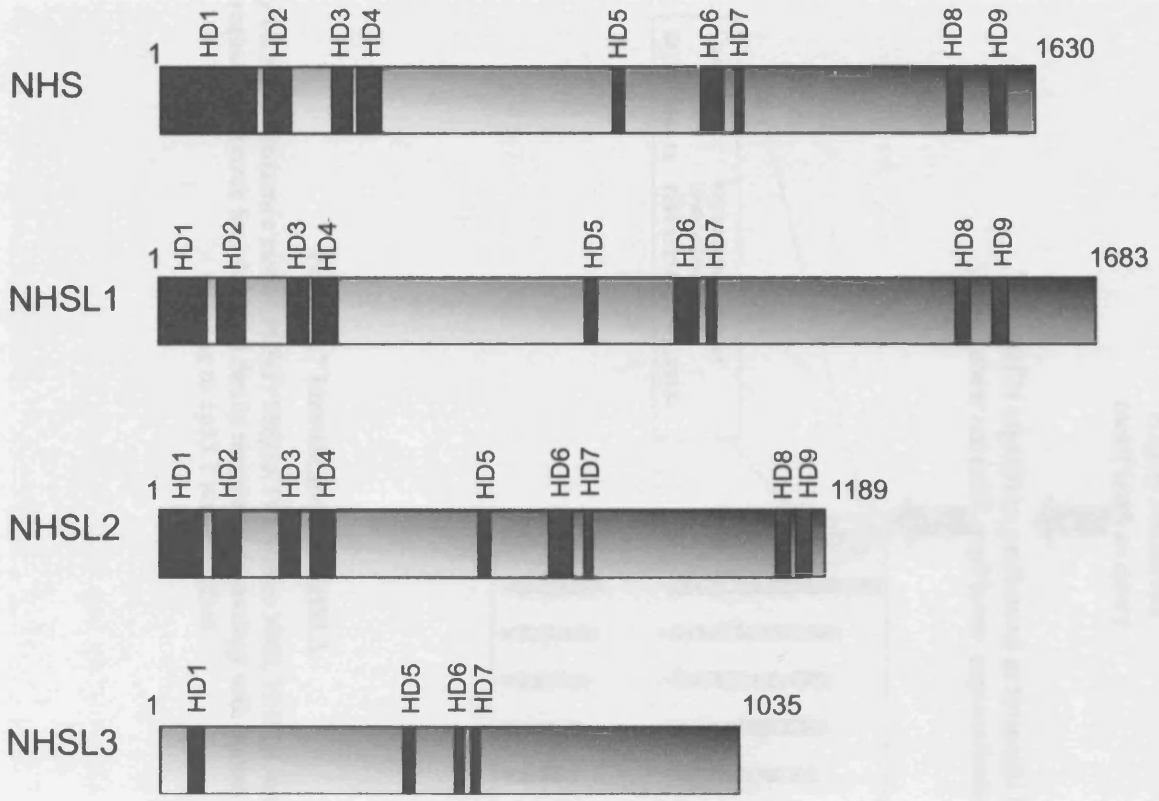


Figure 5.6 Schematic representation of the regions of homology between the NHS protein family.

Positions of homology domains 1-9 (HD1-9) within NHS, NHSL1, and NHSL2 are shown. NHSL3 was found to only contain parts of homology domains 1, 5, 6 and 7. Note, homology domain 9 in NHS and NHSL1 was found to reside in the exact same location. Please refer to figures 5.5 and 5.10 for more detail of the homology domains between the members of the NHS protein family.

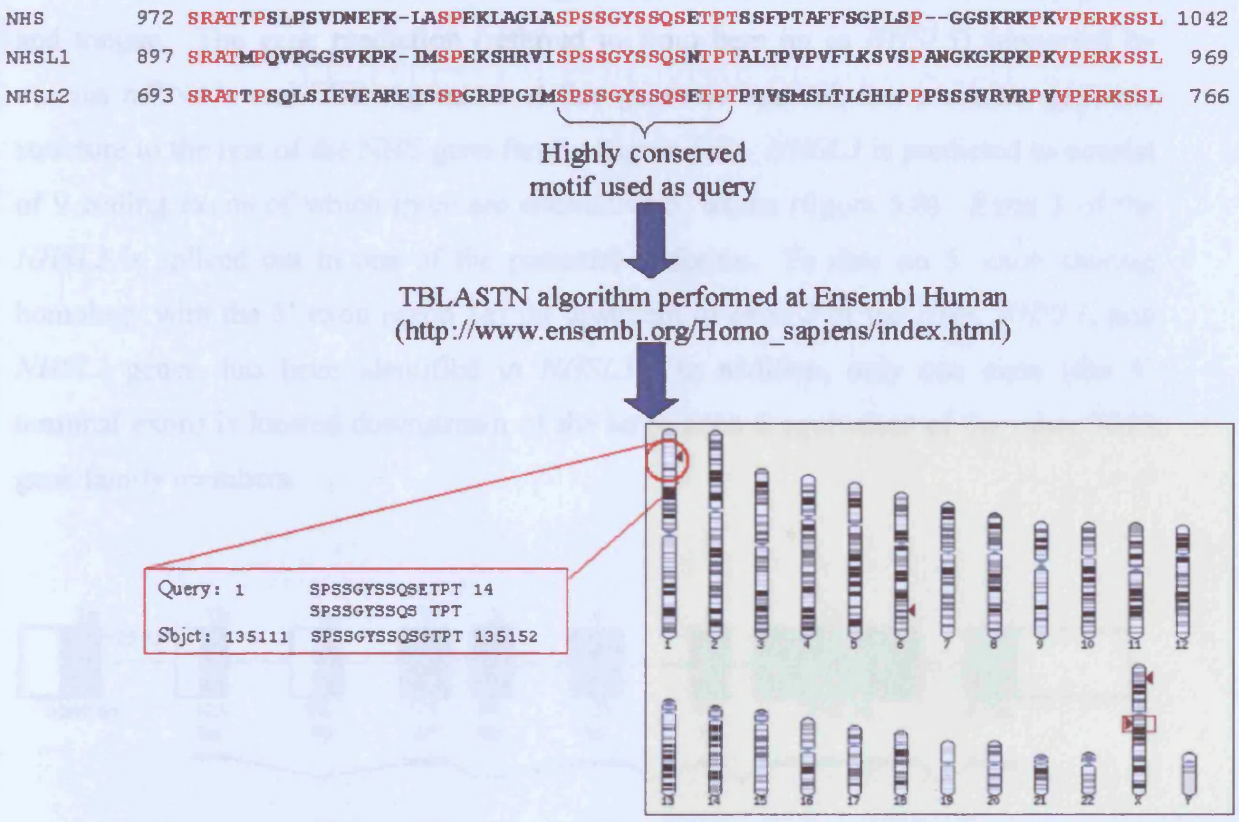


Figure 5.7 Identification of NHSL3.

The highly conserved sequence motif SPSSGYSSQSXTPT within NHS, NHSL1, and NHSL2 was used as query sequence to search for additional family members. Homology with human mRNA AB040955 mapping to 1p35.1 was identified.

Human mRNA AB040955 contains a complete open reading frame spanning 7 coding exons and encoding a 1035 amino acid protein. AB040955 forms part of UniGene cluster Hs.528659, containing numerous mRNA and EST sequences derived from various tissues such as the eye, brain, lung, colon, kidney, liver, heart, mammary gland, and tongue. The gene prediction (referred to from here on as *NHSL3*) supported by various mRNA's and EST sequences at this genomic interval, has a similar genomic structure to the rest of the NHS gene family (figure 5.8). *NHSL3* is predicted to consist of 9 coding exons of which three are alternative 5' exons (figure 5.8). Exon 3 of the *NHSL3* is spliced out in one of the potential isoforms. To date no 5' exon sharing homology with the 5' exon (exon 1a) far upstream of exon 2 of the *NHS*, *NHSL1*, and *NHSL2* genes, has been identified in *NHSL3*. In addition, only one exon (the 3' terminal exon) is located downstream of the large exon 6 equivalent of the other NHS gene family members.

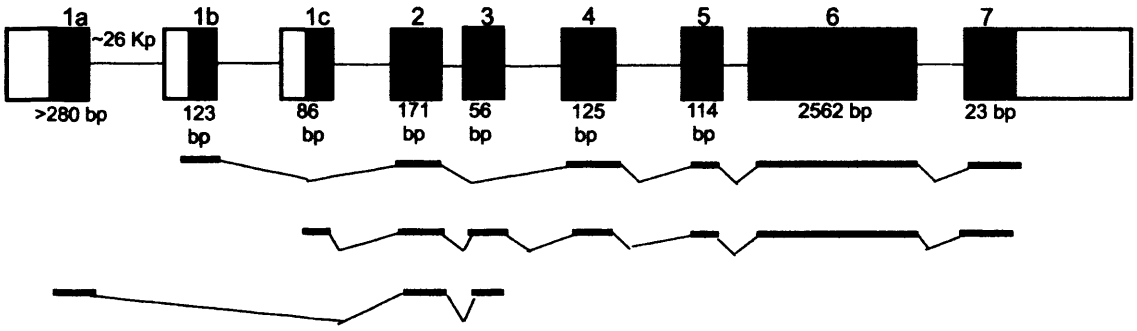


Figure 5.8 *NHSL3* genomic structure.

Blackened squares indicate the open reading frame (ORF), with sizes of coding portions given below. Genomic structure not to scale.

The genomic structures and their chromosomal locations are summarised in figure 5.9.

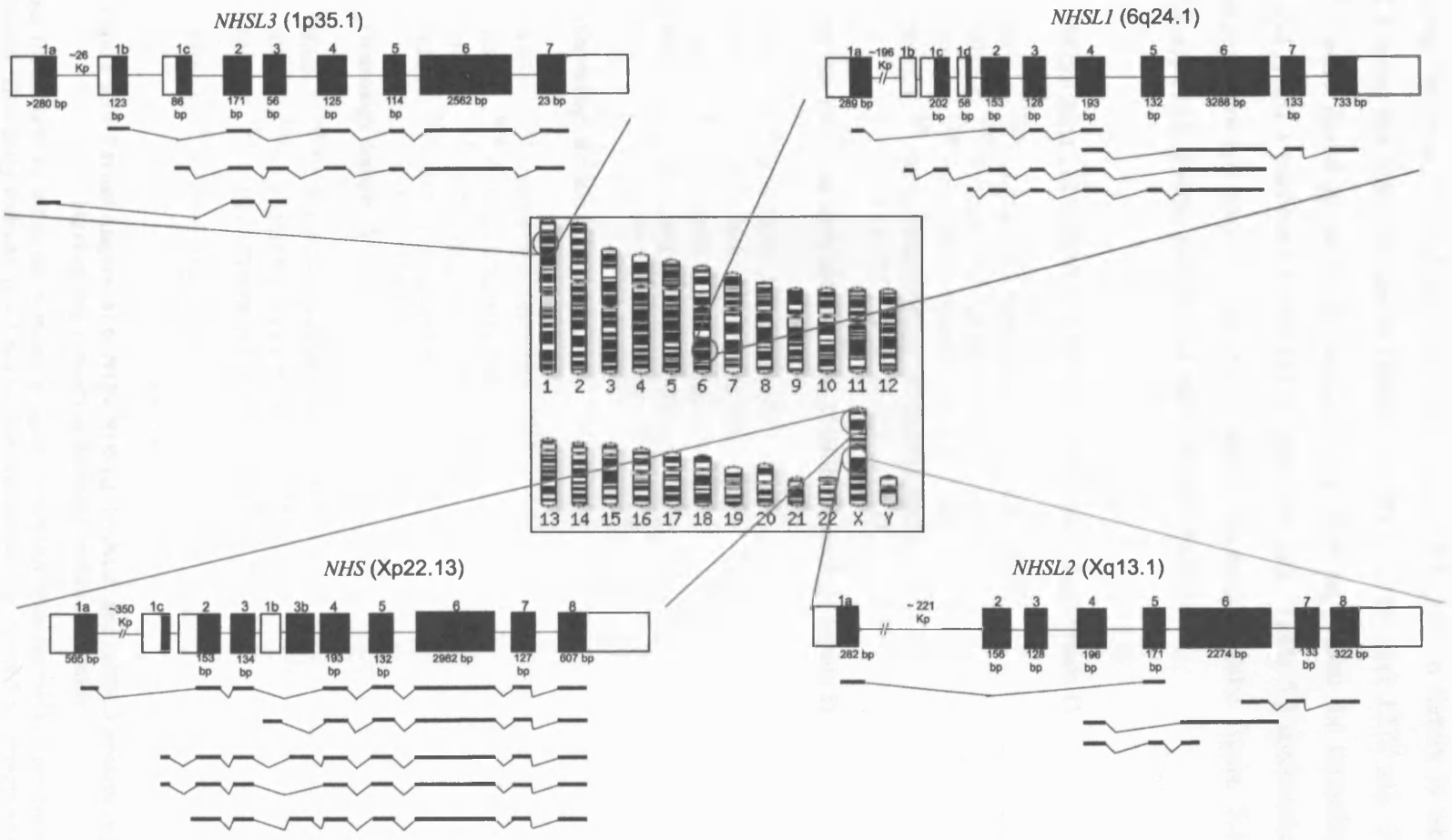


Figure 5.9 Genomic structure, chromosomal localisation, and alternative transcripts of members of the NHS gene family.

Genomic structures not to scale.

Homology between all members of the NHS gene family identified (NHS, NHSL1, NHSL2, and NHSL3) at the protein level is shown in the alignment in figure 5.10. Homology between NHSL3 and the rest of the NHS protein family is weakest, with NHSL3 being the most similar to NHSL1 (22 % identity over 1272 aa). NHSL2 and NHSL3 were found to be 23 % identical over 894 aa, whilst the homology between NHS and NHSL3 was even lower (23 % over 309 aa). Table 5.2 summarises the level of identity between each of the NHS family proteins, whilst figure 5.11 shows a preliminary phylogenetic tree for the NHS protein family.

Within Exon 2 of NHS/NHSL1/NHSL2/NHSL3 (homology domain 1)

```
NHS     212 NIFLPATRPPCVEELHRHARQSLQALRR 239
NHSL1   120 NVFLPTTRPPCVEDLHRQAKLNLKSVLR 147
NHSL2   118 NVFLSSGRPPSVEELLREAQLNLQSLLQ 145
NHSL3    48 NVFFPSGRPPHLEELHTQAQEGLRSLQH  75
           *:*:.: *** :*:*  .*: .*::: :
```

Within Exon 6 of NHS/NHSL1/NHSL2/NHSL3 (homology domain 5)

```
NHS     818 SISFRKPKAKPTPPKRSSSL 837
NHSL1   739 NISLKKAKKPPLPPSRTDSL 758
NHSL2   571 SISLRKAKKKPSPPTRSVSL 590
NHSL3   429 SVSLRKLKRPPPPPRRTHSL 448
            .:*::* *  * ** *: **
```

(homology domain 6)

```
NHS     999 SPSSGYSSQSETPTSS 1014
NHSL1   924 SPSSGYSSQSNTPTAL 939
NHSL2   723 SPSSGYSSQSETPTPT 738
NHSL3   520 SPSSGYSSQSGTPTLP 535
            ********** ***
```

(homology domain 7)

```
NHS    1026 PGGS--KRKPKVPERKSSL 1042
NHSL1   951 PANGKGKPKPKVPERKSSL 969
NHSL2   750 PPSSSVRVRPVVPERKSSL 768
NHSL3   546 P----GKAQPPKPERVTSL 560
            *      : :*  *** :**
```

Figure 5.10 Partial Alignment of NHS, NHSL1, NHSL2, and NHSL3 protein sequences highlighting conserved domains within the family.

Note, the majority of conserved domains in figure 5.5 are lost when the NHSL3 protein sequence is included. Homology domains 1, 5-7 can be seen schematically in figure 5.6. Protein sequences for NHSL1, NHSL2, and NHSL3 are shown in appendix A.

Protein	NHS	NHSL1	NHSL2	NHSL3
NHS	100 %	28 % (474/1694)	25 % (364/1454)	23 % (71/309)
NHSL1	28 % (474/1694)	100 %	23 % (368/1599)	22 % (280/1272)
NHSL2	25 % (364/1454)	23 % (368/1599)	100 %	23 % (206/894)
NHSL3	23 % (71/309)	22 % (280/1272)	23 % (206/894)	100 %

Table 5.2 Summary of the level of identity between each of the NHS protein family members.
The number of identical residues over a given stretch of amino acids is shown in brackets.

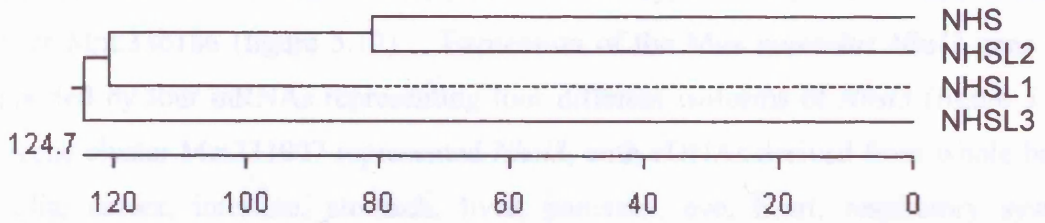


Figure 5.11 A phylogenetic tree of the NHS protein family
The length of each pair of branches represents the distance between sequence pairs, while the units at the bottom of the tree indicate the number of substitution events. The dotted line on the phenogram indicates a negative branch length, a common product of averaging

Following the identification of the NHS family of proteins in Homo *sapiens*, the possibility that *NHSL1*, *NHSL2*, and *NHSL3* may also reside within other species was investigated. In this study, the genome of Mus *musculus* was analysed for potential NHSL1, NHSL2, and NHSL3 orthologs. Using each of the protein sequences for NHSL1, NHSL2, and NHSL3 in Homo *sapiens* as query, BLAST was performed. Perhaps not surprisingly, orthologs for each NHSL1, NHSL2, and NHSL3 were identified in Mus *musculus*. An *NHSL1*, *NHSL2*, and *NHSL3* ortholog were identified on chomosomes 10qA3, XqC3, and 4qD2.2, respectively. Each of these chromosomal regions are syntenic with the corresponding Homo *sapiens* chromosomal bands in which the *NHSL1*, *NHSL2*, and *NHSL3* genes reside. For each of the Mus *musculus* *NHSL1*, *NHSL2*, and *NHSL3* genes, a number of mRNAs were found in the database supporting the expression of each of these genes. Unigene cluster Mm. 158390 supported the expression of *NHSL1*, with mRNAs derived from embryonic cerebellum, eye, thymus, kidney, and mammary gland. Exons 1, 1b, 2-4, and 6-8 of the *Nhsl1* gene were found to be expressed (figure 5.12).

Expression of exons 1, 1b, 1c, and 2-8 of the *Nhsl2* gene were confirmed by mRNA sequences from whole brain, liver, pancreas, kidney, and ear, and formed UniGene cluster Mm.336186 (figure 5.12). Expression of the Mus *musculus* *Nhsl3* gene was supported by four mRNAs representing four different isoforms of *Nhsl3* (figure 5.12). Unigene cluster Mm331907 represented *Nhsl3*, with cDNAs derived from whole brain, ganglia, cortex, intestine, stomach, liver, pancreas, eye, heart, respiratory system, thymus, mammary gland, muscle, bladder, kidney, skin, endocrine system, and ear. Translation begins in exon 4 of isoform 1, and in exons 1b, 1c, and 1d of isoforms 2, 3, and 4, respectively (figure 5.12). Sizes of exons 2, 3, 5, and 7 are identical between Homo *sapiens* and Mus *musculus*. The genomic structures for the Mus *musculus* *Nhsl1*, *Nhsl2*, and *Nhsl3* genes are represented in figure 5.12, whist an phylogenetic tree summarising orthologous and paralogous relationships between the NHS protein family, is shown in figure 5.13.

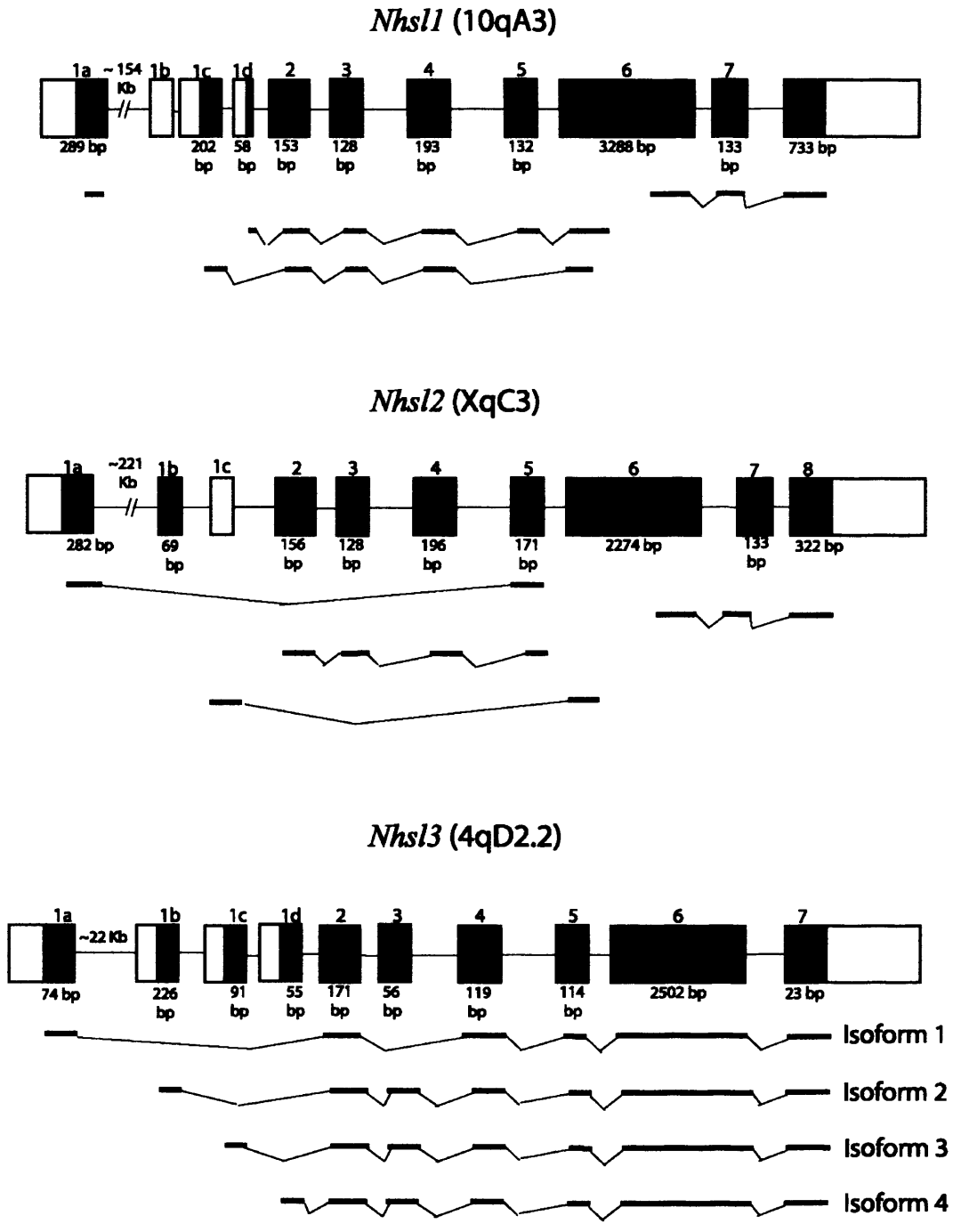


Figure 5.12 Predicted genomic structures of the *Nhsl1*, *Nhsl2*, and *Nhsl3* genes in Mus *musculus*.

Exons are numbered with respect to the corresponding human genes. EST/cDNA hits are depicted below the corresponding genomic structures. Genomic structures not to scale.

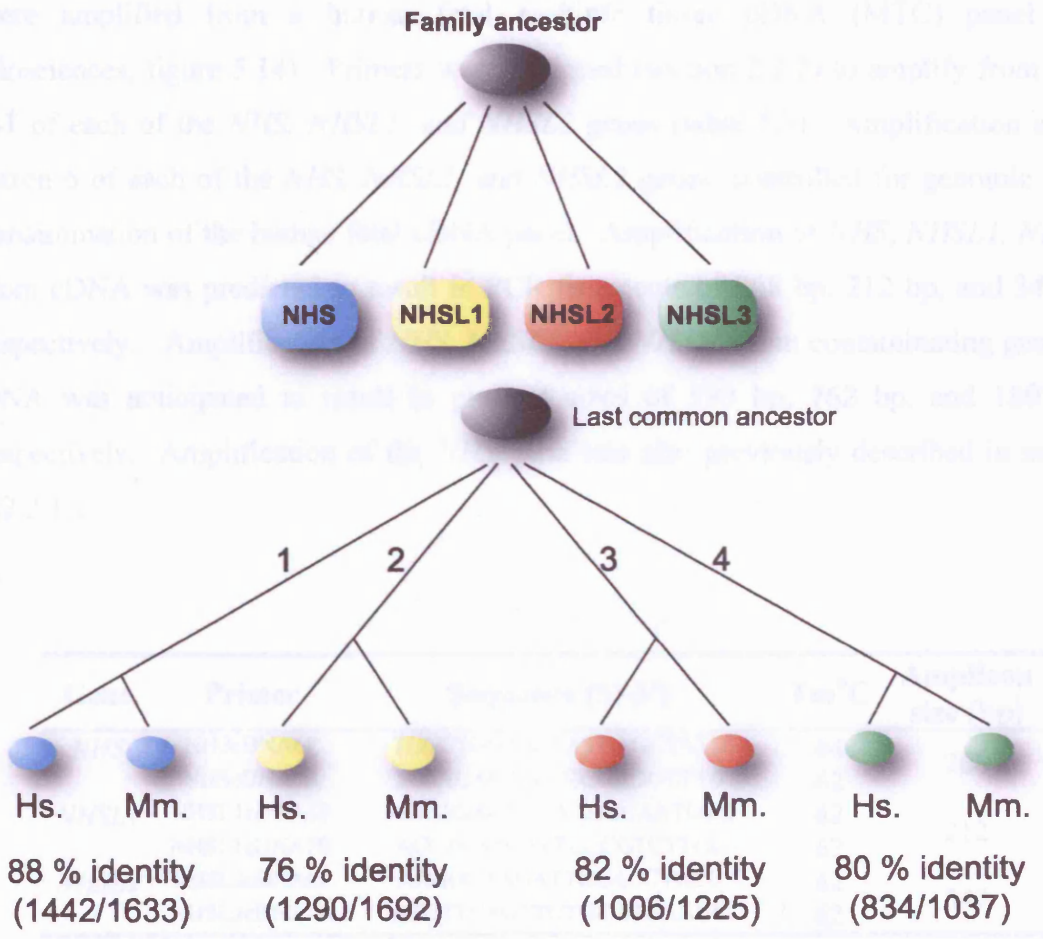


Figure 5.13 A phylogenetic tree illustrating the orthologous and paralogous relationship between the NHS family of proteins.

Each of the four branches denotes the descendants of the four ancestral genes (NHS, NHSL1, NHSL2, and NHSL3) in Homo *sapiens* (Hs.) and Mus *musculus* (Mm.). Percentage identity between orthologs is indicated below each of the four branches, with the number of identical residues over a given stretch of amino acids shown in brackets.

5.2.2.1 Expression of the NHS gene family

To confirm and compare expression of the NHS family of genes (with the exception of *NHSL3*, which was only recently identified), exons 6-8 of *NHS*, *NHSL1*, and *NHSL2* were amplified from a human fetal multiple tissue cDNA (MTC) panel (BD Biosciences, figure 5.14). Primers were designed (section 2.2.2) to amplify from exon 6-7 of each of the *NHS, NHSL1,* and *NHSL2* genes (table 5.3). Amplification across intron 6 of each of the *NHS, NHSL1,* and *NHSL2* genes, controlled for genomic DNA contamination of the human fetal cDNA panel. Amplification of *NHS, NHSL1, NHSL2* from cDNA was predicted to result in PCR fragments of 268 bp, 212 bp, and 347 bp, respectively. Amplification of *NHS, NHSL1,* and *NHSL2* from contaminating genomic DNA was anticipated to result in product sizes of 590 bp, 762 bp, and 1807 bp, respectively. Amplification of the *NHS* gene was also previously described in section 4.2.2.1.2.

Gene	Primer	Sequence (5'-3')	Tm°C	Amplicon size (bp)
NHS	NHScDNA6F	TGCCTGGTACTATCAGCTATG	64	268
	NHScDNA7R	TCCTCAGTTGTTCGAGGTTTG	62	
NHSL1	NHSL1cDNA6F	AGCAGAGTGGAAGCCAATGTC	62	212
	NHSL1cDNA7R	ACCTCATCATTCCCGTCTTCC	62	
NHSL2	NHSL2cDNA6F	AAGGCAAGATTCCACCTCCC	62	347
	NHSL2cDNA7R	ATCTTCAGTTGTGCGTGAAGC	62	

Table 5.3 Primer sequences used to amplify the *NHS, NHSL1,* and *NHSL2* genes in a human fetal cDNA panel.

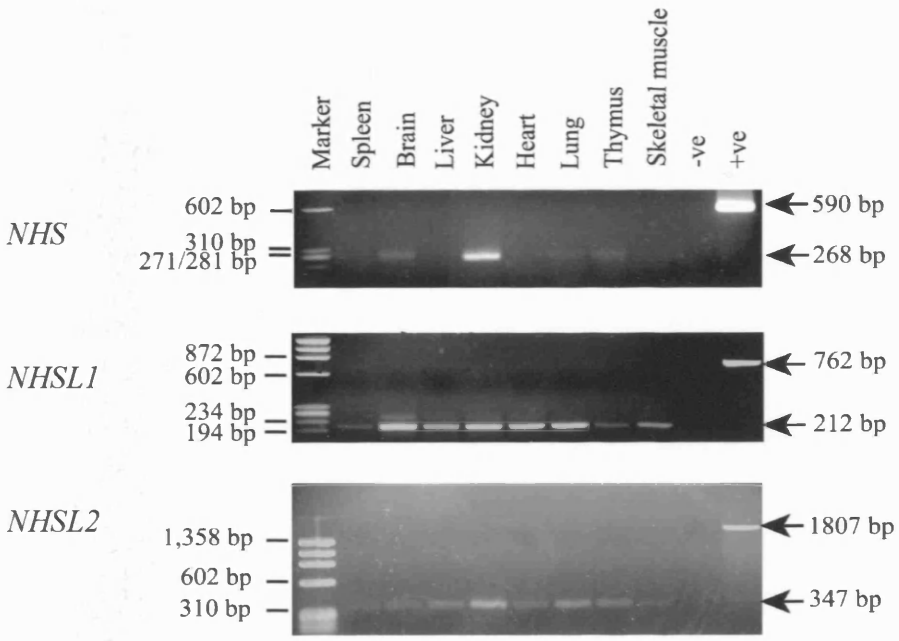


Figure 5.14 Expression analysis of *NHS*, *NHSL1*, and *NHSL2* in a human fetal cDNA panel.

Primers were designed to amplify exon 6 – 8 of each of the *NHS*, *NHSL1*, and *NHSL2* genes; +ve, genomic DNA control; -ve, no DNA control. Products were resolved on a 2 % (w/v) agarose gel.

As described previously (section 4.2.2.1.2), *NHS* was found to be expressed in human fetal brain, thymus, lung and kidney. *NHSL1* and *NHSL2* appear to be much more widely expressed and were detected in all 8 tissues of the human fetal multiple tissue cDNA (MTC) panel (figure 5.14).

5.2.3 Discussion

NHS orthologs were identified through bioinformatic analysis, in Canis *familaris*, Mus *musculus*, Rattus *norvegicus*, Pan *Troglodytes*, Pongo *pygmaeus*, Gallus *gallus*, Bos *Taurus*, Xenopus *laevis*, Danio *rerio*, and Tetradon *nigroviridis*. The NHS protein is highly conserved across species (87-95 %) for which a full length Homo *sapiens* NHS isoform A, protein sequence was determined (Canis *familaris*, Mus *musculus*, Rattus *norvegicus*, and Pan *Troglodytes*). The level of conservation between NHS and the partial orthologous sequences identified in Pongo *pygmaeus*, Gallus *gallus*, Bos *Taurus*, Xenopus *laevis*, Danio *rerio*, and Tetradon *nigroviridis*, ranges from 39-99 %.

The genomic structure for all the predicted full length NHS orthologs (Canis *familaris*, Mus *musculus*, Rattus *norvegicus*, and Pan *Troglodytes*) was deduced and found to be conserved (figure 5.2). Each of these orthologs (except for Pan *Troglodytes*, which is currently missing exon 5) have been found to contain all of the 8 coding exons of *NHS* isoform A (figure 4.9), although expression data to support these predicted exons in the different species is only partially available for Mus *musculus*. However, some of the available mRNA sequences for Mus *musculus* does support the presence of an Exon 3b, which forms part of isoforms C and E of the *NHS* gene. Exon sizes for the 5 orthologs (Homo *sapiens*, Canis *familaris*, Mus *musculus*, Rattus *norvegicus*, and Pan *Troglodytes*) are identical for exons 2, 3, 4, 5, and 7 (except for exons 2 and 5 of Pan *Troglodytes*). An exon 5 equivalent in Pan *Troglodytes* has not yet been identified, which may be a result of incomplete genomic sequence. Further, exon 2 appears to be larger compared to its orthologs. The size of exon 2 in Pan *Troglodytes* is most likely to be an error in the genomic sequence as exon 2 was found to be the most highly conserved exon across species at the protein level. The large (over 2.9 kb) exon 6 exhibits the greatest variability in size and sequence amongst species (figures 5.1 and

5.2), although the size of exon 6 is identical between primates (Homo *sapiens* and Pan *Troglodytes*) and between rodents (Mus *musculus* and Rattus *norvegicus*). Since the genomic structures of Canis *familaris*, Mus *musculus*, Rattus *norvegicus*, and Pan *Troglodytes* are only predictions with little or no supporting mRNAs or ESTs, individual exon sizes may prove to be identical across all species with the possible exception of the largest exon, exon 6.

Together, the expression data presented in chapter 4 of this thesis (detailing the presence of *NHS* in fetal tissue) and the conservation of the NHS protein across species, supports a role for the NHS protein in development.

The most interesting *NHS* ortholog identified to date, and the only ortholog for which there is EST and mRNA data to support its expression, is the mouse (Mus *musculus*) Nhs protein. The Xcat mouse is believed to be a disease model for either CXN and/or NHS due to the similarities in the lens phenotype between CXN, NHS and Xcat. Further, the Xcat mutation has been mapped to an interval on the mouse X chromosome syntenic with Xp22 (to which the CXN and NHS loci are mapped). In light of this, the mouse (Mus *musculus*) *Nhs* gene is an excellent candidate gene for Xcat. Full characterisation of the mouse *Nhs* gene may elucidate additional exons not currently identified in the human NHS gene. Such newly identified exons could then be screened in the CXN family and any NHS families for which a mutation in the *NHS* gene has not been identified. An advantage of looking for disease genes in mouse models is the possibility of using tissue to compare the level of expression of candidate genes between, for example, control and Xcat mice. If the Xcat mouse results from a mutation in some regulatory sequence of the *Nhs* gene, then notable differences in the expression of *Nhs* between Xcat and control mice should be detected.

Following the identification of *NHS* orthologs, the human *NHS* gene was found to form a new gene family through the discovery of three *NHS* paralogs, *NHSL1*, *NHSL2*, and *NHSL3*. Residing on chromosome 6q24.1, *NHSL1* was the first of the three paralogs identified (figure 5.9). NHS shares 28 % protein sequence identity with the predicted NHSL1 protein, 25 % sequence identity with NHSL2, and 23 % with NHSL3 (table 5.2). The genomic structure of *NHS, NHSL1, NHSL2,* and *NHSL3* are highly similar,

providing further support that *NHSL1*, *NHSL2*, and *NHSL3* are paralogs of the *NHS* gene.

Initial pile up analysis of the NHS, NHSL1, and NHSL2 protein sequences highlighted conserved domains across the NHS family. One such conserved domain (sequence motif, SPSSGYSSQSXTPT) led to the identification of the NHSL3 gene (figure 5.7). A second alignment of all four protein sequences (NHS, NHSL1, NHSL2, and NHSL3), highlighted fewer conserved domains amongst the NHS protein family, with the SPSSGYSSQSXTPT motif being the most highly conserved (only amino acid designated X differs between each of the four members of the NHS protein family). Therefore, the sequence motif SPSSGYSSQSXTPT must play a crucial role in the function of the NHS protein family. Initial bioinformatic analysis of this sequence motif (SPSSGYSSQSXTPT) has not provided any clues as to its potential function. Other conserved domains amongst NHS, NHSL1, and NHSL2 (figure 5.5), may also be of significant importance to the function of this protein family and should be investigated further.

Initial phylogenetic analysis of the NHS family of proteins revealed NHS and NHSL2 to be the most closely related (figure 5.11). Both NHS and NHSL2 are located on the X chromosome at Xp22.13 and Xq13.1, respectively. NHSL1 appears to have resulted from a duplication event prior to the duplication event of NHS, which resulted in NHSL2.

Interestingly, the *NHS* and *NHSL1* genes lie on paralogous duplicated chromosomal intervals on Xp22 and 6q24, highlighted by the presence of the paralogs *REPS1* and *REPS2*, which reside telomeric to *NHS* and *NHSL1* respectively. Thus, the chromosomal region on Xp22 harbouring *NHS* and *REPS1*, arose from duplication of the chromosomal region 6q24 (containing *NHSL1* and *REPS2*), or vice versa. The NHSL3 protein and its gene on chromosome 1p35.1 is the most distantly related of the NHS family.

Based on preliminary investigation into the NHS gene family, it would appear that *NHS*, *NHSL1*, *NHSL2*, and *NHSL3* are not confined to Homo *sapiens*. As one might expect, all four members of the NHS gene family were also identified in Mus *musculus*. It is

clear therefore, that the NHS gene family did not arise solely from gene duplication in the Homo *sapiens* lineage following divergence of primates and rodents. Thus, the common ancestral gene of the NHS gene family existed prior to the last common ancestor of primates and rodents. Duplication of the common ancestral gene for the NHS gene family therefore resulted in paralogous genes prior to the last common ancestor of rodents and primates, which went on to become progenitors of the primate and rodent lineages. Subsequently, each NHS gene family member in Homo *sapiens* and Mus *musculus* is a paralog of each of the remaining NHS gene family members in both species. For example, the *NHS* gene in Homo *sapiens* is a paralog of NHSL1, NHSL2, and NHSL3 in both Homo *sapiens* and Mus *musculus*, and is an ortholog of the *NHS* gene in Mus *musculus*. Since only four NHS gene family members (*NHS*, *NHSL1*, *NHSL2*, and *NHSL3*) have been identified in Homo *sapiens* and Mus *musculus*, it would appear that evolution from the last common ancestor of both lineages has been of vertical inheritance only.

Expression analysis of the *NHS*, *NHSL1*, and *NHSL2* in a human fetal cDNA panel, revealed *NHSL1* and *NHSL2* to be more widely expressed (transcripts were detected in all eight of the fetal tissues tested), whilst expression of *NHS* was only detected in fetal brain, kidney, lung, and thymus.

Combining the preliminary expression data of *NHS*, *NHSL1*, and *NHSL2* in this thesis with the knowledge that each of these proteins arose from gene duplication prior to the common ancestor of Homo *sapiens* and Mus *musculus* and that the sequence is well conserved between these species, all members of the NHS family of proteins are likely to play key roles in development. This is likely for *NHSL3*, as available mRNA data supporting its expression are derived mainly from embryonic tissues.

The rodent lineage split 12 to 24 million years ago into separate lines that gave rise to the rat and mouse, whereas primates and rodents diverged around 80 million years ago. Since all four NHS family members were identified in humans and mice, all four members are likely to be present in rat and other primates such as the chimpanzee. It will be important to determine how far back in evolutionary terms, the four members of the NHS protein family (*NHS*, *NHSL1*, *NHSL2*, and *NHSL3*) go. For example, the WAVE (WASP family verprolin homologous) proteins involved in regulating actin

dynamics, are present as 3 members (WAVE1-3) in mammals. In Drosophila however, only one WAVE homolog (Scar) is currently known (Bear *et al.*, 1998). Paralogs generally perform distinct biological functions, albeit through the same mechanism. Understanding in more detail the mechanisms through which the NHS protein acts may shed light on the roles of the remaining NHS family members.

- Chapter Six -

Functional analyses of the NHS protein

6.1 Introduction

To date, five potential transcripts for the human *NHS* gene have been identified from databases (figure 4.9), which are predicted to encode proteins of 1,630 aa (isoform A), 1,335 aa (isoform B), 1,474 aa (isoform C), 1,453 aa (isoform D), and 1,473 aa (isoform E). All isoforms differ in the 5' exon except for isoforms C and D, which differ by the alternative splicing of exon 3b (present in isoform C). Apart from the differing 5' exon between the isoforms, all (with the exception of isoform B), contain exons 2-8 of the *NHS* gene (figure 4.9). Translation of the NHS protein in isoform B has been predicted to start in exon 4 at the first methione and continue through exons 5-8 (Burdon *et al.*, 2003). It should be noted, however, that there are currently no published experimental data demonstrating protein expression of any isoform in any tissue or cell type.

Burdon and colleagues reported no significant homology between the NHS protein and any other protein or class of proteins. The only sequence motifs identified in NHS (Burdon *et al.*, 2003) were four potential monopartite nuclear localisation signals (NLS) at amino acid positions: 371-379 (RRRKLRRRK), 438-444 (PSRRRIR), 822-825 (RKPK), 1026-1034 (PGGSKRKPK).

As the function of the NHS protein is currently unknown, homologies and functional clues were sought through bioinformatic analysis as detailed in this chapter. A powerful approach to identifying any potentially significant domains came from

identifying the *NHS* paralogs, *NHSL1*, and *NHSL2* (section 5.2.2). Alignment of the protein sequences of NHS, NHSL1, and NHSL2 (figure 5.5) highlighted conserved domains, which were used to perform BLAST algorithms against the NCBI database. Importantly, this resulted in the identification of sequence similarity at the protein level of NHS, NHSL1, and NHSL2 with the WAVE homology domain, WHD (also referred to as the SCAR homology domain, SHD; figure 6.1).

This chapter describes protein similarities identified and strategies carried out to address their significance. An anti-NHS peptide antibody was raised and characterised for use in immunocytochemistry experiments and competitive actin polymerisation assays were conducted to test the potential function of identified domains within the NHS protein.

6.2 Results

Monopartite NLS motifs are characterised by a cluster of around five basic residues. The putative monopartite nuclear signals identified in NHS (Burdon *et al.*, 2003) are conserved across species (this thesis, figure 5.1) and the monopartite motif residing at amino acid position 822-825 in NHS is quite well conserved between NHS, NHSL1, and NHSL2 (this thesis, figure 5.5).

Bipartite nuclear localisation signals are characterised by the presence of two adjacent basic amino acids (Arg or Lys), a spacer of any 10 residues followed by at least three basic residues in the five positions after the spacer region. In this thesis, analysis of the NHSL1, NHSL2, and NHSL3 proteins identified potential bipartite nuclear localisation signals at amino acid positions 265-281 (NHSL1, **RR**SLIYTDTLVR**R**P**KK**V), 63-79 and 54-75 (NHSL2, **RR**TDSLYRRTVRL**RRR**L and **RR**QRSKSISLRKA**KKK**P, respectively; see appendix A). No bipartite nuclear localisation signals have been identified in NHSL3 at present. Predicted amino acid sequences for NHSL1, NHSL2, and NHSL3 are shown in appendix A.

6.2.1 Homology to the Wiskott Aldrich Syndrome (WASP) family of proteins

6.2.1.1 The WAVE homology domain, WHD

Through *in silico* analysis, identity shared between the N-termini of NHS, NHSL1, and NHSL2 (coded for by exons 1a-2) and the WHD of the WAVE proteins, was identified by BLAST analysis using the amino acid sequences of NHS, NHSL1, and NHSL2 as query. Within the WHD the NHS protein contains an intervening amino acid sequence (amino acids 22-119, figure 6.1, table 6.1) coded by exon 1, with no similarity to the WHD, which contains a number of proline residues (25 % proline). The potential importance of the putative WHD in the NHS protein is highlighted not only by the conservation between the paralogs but also the N-terminus of NHS is highly conserved amongst its orthologs (74 % identity from amino acids 1-219; figure 5.1).

The high proline content (amino acids 22-119) of the NHS protein is present in all full-length NHS protein orthologs identified (figure 5.1). The level of conservation over this stretch of amino acids (22-119) however, is lower when compared with the remainder of the putative NHS WHD (amino acids 1-21 and 119-219, figure 5.1). The high proline content is not present at the N-terminus NHSL1 or NHSL2, although NHSL1 does contain a proline-rich region (38 % proline over 145 amino acids, see table 6.1) in the centre of its full-length protein sequence. To date, no proline-rich region has been identified in NHSL2.

Proteins known to contain a WHD also have proline-rich regions. The WHD has been described in the WAVE (WASP family Verprolin homology protein; Suetsugu *et al.*, 1999) family of proteins (also called SCAR and WASF, section 1.11). WAVE proteins 1, 2 and 3 are part of a larger family of proteins called the WASP family, which also includes WASP and N-WASP, as discussed in chapter 1.11. Members of this family have a conserved C-terminal region consisting of a proline-rich segment, one or two WASP homology 2 domains (WH2, also known as Verprolin homology

Family	Protein	Size (a.a.)	Domain and position								
			WH1	Basic	GBD	Proline rich	WH2	Central/Acidic	WHD	Alanine rich	Serine rich
WASP	WASP	502	38-144	225-249	237-250	312-421	431-447	465-502	-	-	-
	N-WASP	505	34-138	165-181	203-216	277-392	405-422 433-450	466-505	-	-	-
	WAVE1	559	-	173-206	-	275-492	497-514	526-559	1-165	-	-
NHS	NHS	1630	-	364-387	-	22-119	1418-1438	545-615 665-704	1-219*	88-133	1440-1629
	NHSL1	1683	-	265-285	-	999-1143	1428-1448	1512-1558	1-127	-	967-991
	NHSL2	1189	-	564-586	-	-	1102-1122	339-407	1-126	-	664-814
	NHSL3	1035^	-	144-169	-	447-1006	-	301-368	-	-	328-440

Table 6.1 Domain organisation of the WASP and NHS protein families.

*Also contains the high proline content within the WHD of NHS at amino acids 22-119 inclusive

domain), a central basic motif and an acidic region (table 6.1 and figure 6.2, section 1.11). The N-terminus of these proteins differ, with WASP and N-WASP both containing a WASP homology 1 (WH1) region and a GTPase binding domain (GDB; figure 6.2). In contrast, WAVEs share a WHD at the N-terminus (table 6.1 and figure 6.2).

It is interesting to note that the protein sequence of NHS, NHSL1 and NHSL2 that share sequence homology with the WHD are expressed only within isoform A of the *NHS* gene (figure 4.9) and most likely within only one isoform of the *NHSL1* and *NHSL2* genes. Isoform A also appears to be the major isoform responsible for disease in Nance-Horan syndrome patients (as discussed in chapter four).

Until recently no function had been attributed to the WHD of the WAVE proteins. New evidence has shown that the WHD can act as a binding domain for the Abelson Interactor (Abi) protein family (Echarri *et al.*, 2004; Leng *et al.*, 2005, section 1.13). The putative WHD in NHS was therefore examined for key amino acid residues which are conserved across species in the Abi-binding domain of the WAVE protein family. Of the ten key amino acid residues spanning the Abi-binding domain of the WAVE protein family (figure 6.1), five are completely conserved amongst the NHS protein family (figure 6.1). The remaining five amino acids are substituted with residues that are highly similar in their biochemical properties (figure 6.1), implying that a potential Abi-binding domain may exist within the putative SHD of the NHS protein family (NHS, NHSL1, and NHSL2).

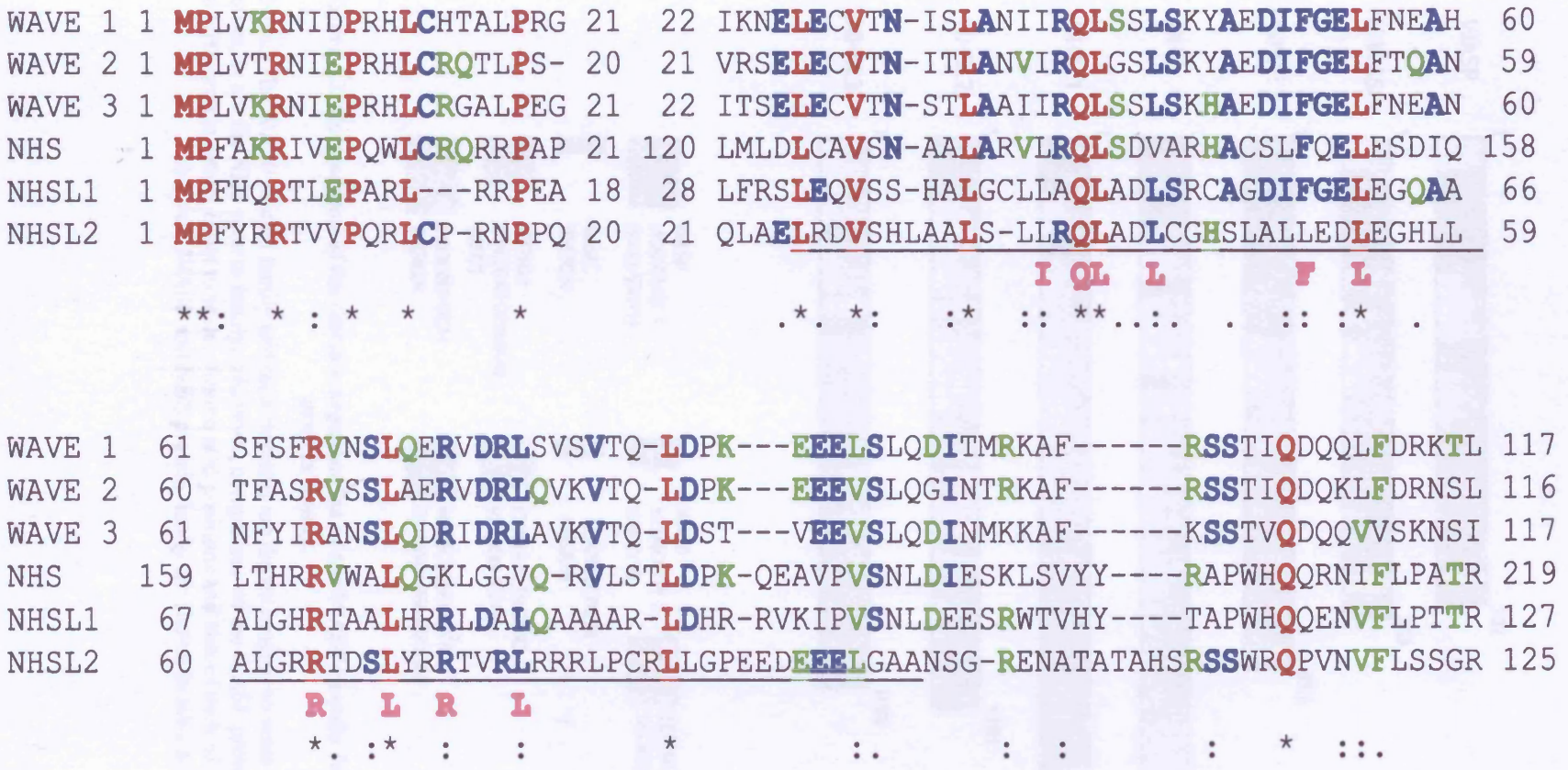


Figure 6.1 Alignment of the N-termini of NHS, NHSL1, and NHSL2 with the WHD of the WAVE protein family.

Amino acids highlighted in red are conserved; blue residues are completely conserved in the WAVEs and in at least one of the NHS family proteins. Green residues indicate amino acids present in at least one WAVE protein and at least one NHS protein family member. Those residues highlighted in pink underneath are conserved in mammalian WAVEs, Drosophila SCAR, and Dictyostelium SCAR. The Abi binding domain is underlined.

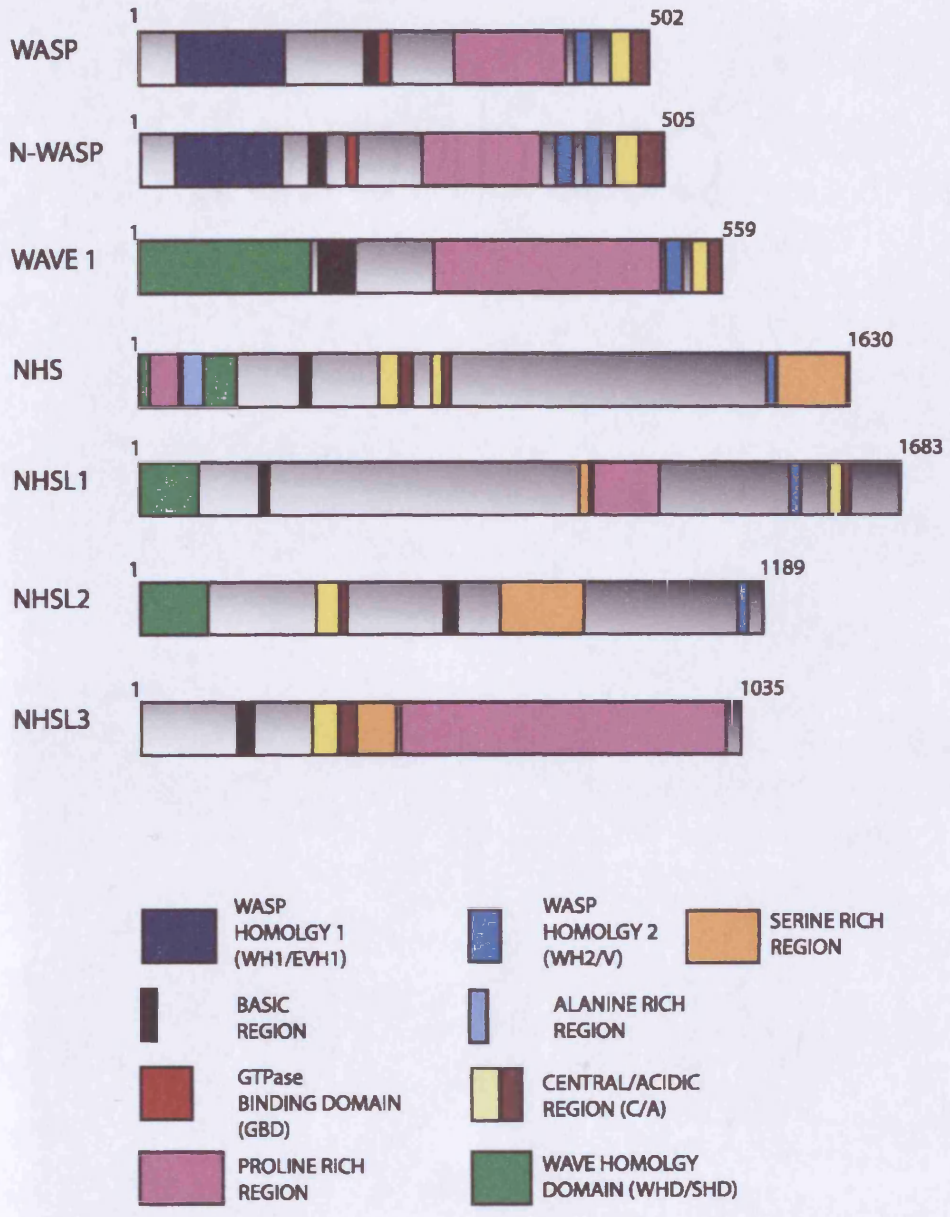


Figure 6.2 Comparison of the domain organisation of the WASP protein family with the NHS protein family.

Note: The WASP protein family and their domains are drawn roughly to scale with respect to one another, as are the NHS protein family. However, comparison of the WASP protein family with that of the NHS protein family is not to scale. Amino acid positions and size of each of the above domains in both the WASP and NHS protein family are shown in table 6.1.

6.2.1.2 The WCA (WASP-homology, central region, acidic region) domain

Identification of homology to the WHD of the WAVE proteins prompted a thorough search for other possible domains shared between the WASP protein family and the NHS family. Using the SIM alignment tool at ExPASy (section 2.7.4), the NHS protein was aligned with the WCA (WASP-homology, central region, acidic region) domain of the various WASP family proteins (section 1.11). A stretch of approximately 40 amino acids of the NHS protein (NHS CA1 amino acids 670-709) was found to align uninterrupted with the central (C) and acidic (A) regions of N-WASP (figure 6.3).

A tryptophan residue in the CA domain that has been reported to be conserved amongst members of the WASP family and is though to be essential for binding the Arp2/3 complex (Marchand *et al.*, 2001, section 1.10), was conserved within this putative NHS-CA1 domain (figure 6.3). The conserved tryptophan residue in NHS is preceded by a number of acidic residues (aspartic and glutamic acid, and asparagine) and several serine residues. Aspartic and glutamic acid residues constitute a large portion of the acidic region preceding the conserved tryptophan residue in proteins of the WASP family. In addition, WAVE proteins also contain a number of serine residues in the acidic domain. The acidic domain appears to be variable in length and content (figure 6.3).

Within the central (C) region of all known Arp2/3 activators (except cortactin) lies a conserved arginine residue (R477 in WASP) that is also conserved within this stretch of 40 amino acids of the NHS protein (figure 6.3). The central region in members of the WASP protein family also contains conserved long-chain aliphatic residues such as Leucine, Isoleucine and Valine (figure 6.3). Together with the conserved arginine residue, these aliphatic residues have been reported to form a sequence motif that interacts with the Arp2/3 complex leading to its activation (Panchal *et al.*, 2003). The putative central region in NHS (NHS CA1 670-688) also contains the conserved leucine residue, which when mutated abolishes the ability of an Arp2/3 activator to activate the Arp2/3 complex (figure 6.3, Panchal *et al.*, 2003).

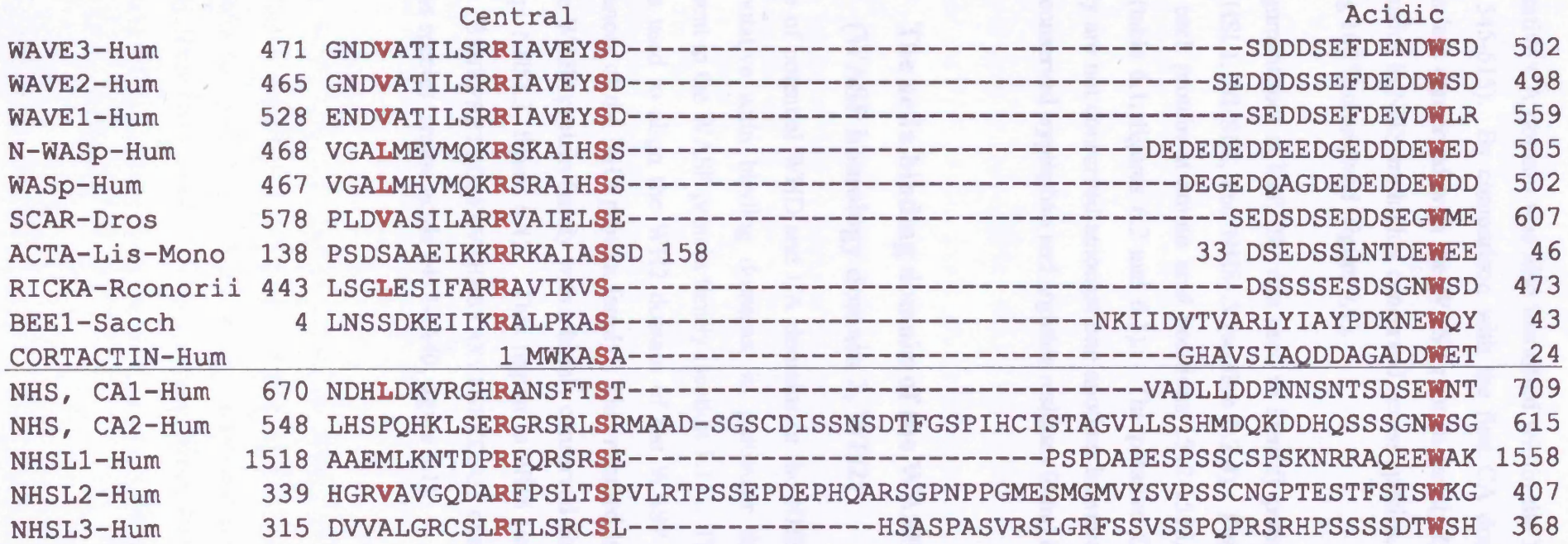


Figure 6.3 Multiple protein alignment of the central and acidic regions (CA) for known activators of the Arp2/3 complex with the NHS protein family

Lis-Mono = Listeria *monocytogenes*, Sacch = Saccaromyces *cerevisiae*, RConorii = Rickettsie *conorii*. Key amino acids within the CA domains, which are highly conserved are highlighted in red.

A second putative CA domain was also identified within the NHS protein (NHS CA2 amino acids 545-615). By comparison with the first CA domain identified in NHS, however, it is less conserved with the WASP protein family CA domains (figure 6.3). Both CA domains in NHS are highly conserved across species, with the tryptophan and arginine being 100 % conserved (figure 5.1).

The SIM alignment tool at ExPASy was used to identify potential central/acidic (CA) regions in NHSL1, NHSL2, and NHSL3 (section 2.7.4). Putative CA domains were identified in each protein at amino acid positions 1512-1558, 339-407, and 301-368, respectively (table 6.1, figures 6.2 and 6.3). The putative CA domains in the NHS protein family are not conserved amongst one another, however all of them do contain the essential conserved tryptophan and arginine residues found in Arp2/3 activators.

6.2.1.3 The actin binding domain of the WASP protein family (WASP homology domain 2, WH2)

The presence of potential WHD and CA domains in the NHS protein family led to a search for putative actin binding domains, in particular the actin-binding (WH2) domains present in the WASP protein family (section 1.11). The SIM alignment tool at ExPASy was used to align the WH2 domain of the WASP protein family with the protein sequences of the NHS protein family. This resulted in alignment of the WH2 domain in the WASP protein family with a highly conserved domain in the NHS protein family (except NHSL3, figure 6.4). This region in NHS lies within a stretch of 29 amino acids (SPNKPRTTEDLFAVIHRSKRKVLGRKDSG) over which there is 90 % identity across species (amino acids 1412-1440, figure 5.1).

```
NWASP WH2a      ALLDQI----RQGIQL-KS
NWASP WH2b      ALLDQI----REGAQL-KK
WASP            ALLDQI----RQGIQL-NK
WAVE1           VLLEAI----RKGIQL-RK
WAVE2           DLLSAI----RQGFQL-RR
WAVE3           DLLAAI----RMGIQL-KK
NHSL1           DLFAAIHRSKRK--VLGRR
NHS             DLFAVIHRSKRK--VLGRK
NHSL2           DLFTVIHRSKRK--LLGWK
```

Figure 6.4 Alignment of the WH2 domain in the WASP protein family with the NHS protein family.

Note, no alignment to NHSL3 could be obtained. Amino acids highlighted in red are completely conserved. Blue residues are found in at least one member of each family (WASP and NHS).

6.2.1.4 Additional domains within the NHS protein family

The homologies identified between the WASP protein family, in particular the WAVE subfamily, and the NHS protein family are striking and imply that the NHS protein family are functionally related to the WAVE subfamily. In addition to the WHD domain, WH2 domain, and the central and acidic domains, the WASP protein family have a basic region at their N-terminus (table 6.1 and figure 6.2, section 1.11). A GTPase binding domain also resides within the protein sequence of WASP and N-WASP, but is absent in the WAVE protein subfamily (figure 6.2, section 1.11). Analysis of the NHS protein family using the ExPASy database (section 2.7.4) of protein families and domains (prosite: http://ca.expasy.org/prosite/) revealed basic regions within each of the NHS family proteins (table 6.1 and figure 6.2). The basic region, like the WAVE subfamily, is located N-terminally in NHS, NHSL1 and NHSL3, and towards the centre of the NHSL2 protein. No potential GTPase binding domain was found in the NHS protein family strengthening the idea that they are more closely related to the WAVE subfamily rather than WASP and N-WASP.

Through the prosite database (section 2.7.4), an alanine rich region was identified within the WHD of the NHS protein (table 6.1 and figure 6.2) but is absent in the rest of the NHS protein family members. Furthermore, a serine rich region (25 % serine) was found in each of the NHS protein family members, which varies in size from 189 aa in NHS to 24 aa in NHSL1 (table 6.1 and figure 6.2). Such serine rich domains are absent from the WASP protein family.

6.2.2 Characterisation of an anti-NHS antibodies

6.2.2.1 Generation of an anti-NHS antibodies

Polyclonal antisera were raised against a synthetic peptide derived from the C-terminus of the NHS protein (figure 6.5), which is present in all NHS isoforms (figure 4.9) and is completely conserved across species (figure 5.1). The very C-terminus of NHS however, is not conserved amongst the NHS protein family. The NHS protein family C-termini differ between all four members. Peptide production, conjugation and generation of polyclonal rabbit anti-NHS antisera was provided commercially by Sigma Genosys, UK. The last 14 amino acids of the NHS protein (figure 6.5) were selected to generate an NHS peptide sequence. A cysteine residue was positioned at the N-terminus of the peptide sequence (CDGSPHDDRSSQSST) for conjugation with keyhole lympet hemocyanin (KLH, a carrier protein). Conjugation was achieved through MBS (3-maleimidobenzoic acid n-hydroxysuccinimide ester) chemistry, whereby the KLH carrier protein is conjugated via the thiol group on the cysteine. Two rabbits (SG1703 and SG1704) were immunised with the synthetic peptide, and the pre-immune and immune antisera provided by Sigma Genosys, UK.

MPFAKRIVEPQWLCRQRRPAPGPAVDASGGSAEPPPPLQPPGRRDLDEVEAPGPEEPARAVPAPSGLPPPPPPLPA
PADQTQPPHGEASVAGEESTAGIPEAAPAAGEASSAAAAAAAVLLMLDLCAVSNAALARVLRQLSDVARHACSLFQE
LESDIQLTHRRVWALQGKLGGVQRVLSTLDPKQEAVPVSNLDIESKLSVYYRAPWHQQRNIFLPATRPPCVEELHR
HARQSLQALRREHRSRSDRREQRAAAPLSIAAPPLPAYPPAHSQRRREFKDRHFLTSHPPEDEDTDVMLGQRPKNP
IHNIPSTLDKQTNWSKALPLPTPEEKMKQDAQVISSCIIPINVTGVGFDREASIRCSLVHSQSVLQRRRKLRRRKT
ISGIPRRVQQEIDSDESPVARERNVIVHTNPDPSNTVNRISGTRDSECQTEDILIAAPSRRRIRAQRGQSIAASLS
HSAGNISALADKGDTMFTPAVSSRTRSRSLPREGNRGGDAEPKVGAKPSAYEEGESFVGDHERTPNDFSEAPSSPS
AQDHQPTLGLACSQHLHSPQHKLSERGRSRLSRMAADSGSCDISSNSDTFGSPIHCISTAGVLLSSHMDQKDDHQS
SSGNWSGSSSTCPSQTSETIPPAASPPLTGSSHCDSELSLNTAPHANEDASVFVTEQYNDHLDKVRGHRANSFTST
VADLLDDPNNSNTSDSEWNYLHHHHDASCRQDFSPERPKADSLGCPSFTSMATYDSFLEKSPSDKADTSSHFSVDT
EGYYTSMHFDCGLKGNKSYVCHYAALGPENGQGVGASPGLPDCAWQDYLDHKRQGRPSISFRKPKAKPTPPKRSSS
LRKSDGNADISEKKEPKISSGQHLPHSSREMKLPLDFANTPSRMENANLPTKQEPSWINQSEQGIKEPQLDASDIP
PFKDEVAESTHYADLWLLNDLKTNDPYRSLSNSSTATGTTVIECIKSPESSESQTSQSESRATTPSLPSVDNEFKL
ASPEKLAGLASPSSGYSSQSETPTSSFPTAFFSGPLSPGGSKRKPKVPERKSSLQQPSLKDGTISLSKDLELPIIP
PTHLDLSALHNVLNKPFHHRHPLHVFTHNKQNTVGETLRSNPPPSLAITPTILKSVNLRSINKSEEVKQKEENNTD
LPYLEESTLTTAALSPSKIRPHTANKSVSRQYSTEDTILSFLDSSAVEMGPDKLHLEKNSTFDVKNRCDPETITSA
GSSLLDSNVTKDQVRTETEPIPENTPTKNCAFPTEGFQRVSAARPNDLDGKIIQYGPGPDETLEQVQKAPSAGLEE
VAQPESVDVITSQSDSPTRATDVSNQFKHQFVMSRHHDKVPGTISYESEITSVNSFPEKCSKQENIASGISAKSAS
DNSKAEETQGNVDEASLKESSPSDDSIISPLSEDSQAEAEGVFVSPNKPRTTEDLFAVIHRSKRKVLGRKDSGDMS
VRSKSRAPLSSSSSSASSITSPSSNVTTPNSQRSPGLIYRNAKKSNTSNEEFKLLLLKKGSRSDSSYRMSATEILK
SPILPKPPGELTAESPQSTDDAHQGSQGAEALSPLSPCSPRVNAEGFSSKSFATSASARVGRSRAPPAASSSRYSV
RCRLYNTPMQAISEGETENSDGSPHDDRSSQSST

Figure 6.5 Amino acid sequence of the NHS protein (isoform A).

Antisera to a C-terminal peptide (NHSC1, highlighted red) was raised in rabbit. The very C-terminus of NHS is present in NHS isoforms A-E and is completely conserved across species.

6.2.2.2 Characterisation of anti-NHS antibody SG1703

In order to characterise the antisera, a GST-NHS fusion protein (NHSC1, figures 6.6 and 6.7) containing the 201 amino acids C-terminal of NHS, was generated. Primers with engineered in frame restriction endonuclease sites were designed (section 2.2.2, table 6.2) to PCR amplify (section 2.2.1) a 607 bp fragment from exon 8 of the *NHS* gene using genomic DNA as template.

Gene	Primer	Sequence (5'-3')	Annealing temperature (°C)	Amplicon size (bp)
NHS	NHSC1F	CG**GGATCC**TCCAAGAGGAAAGTACTTG	63	607
	NHSC1R	CG**GAATTC**TCTATGTTGAACTCTGGGAG	63	
	NHSC1Fseq	TCCAAGAGGAAAGTACTTG	55	
	NHSC1Rseq	TCTATGTTGAACTCTGGGAG	55	

Table 6.2 Primer sequences used to amplify a 607 bp fragment from exon 8 of the *NHS* gene. Restriction endonuclease sites engineered into the primer sequences are highlighted in bold and underlined (GGATCC – BamHI, GAATTC – EcoRI).

After amplification of the 607 bp fragment, PCR products were double digested with BamHI and EcoRI (section 2.2.11) and ligated (section 2.3.1) into the pGEX-2T vector (pre-digested with BamHI and EcoRI).

Two clones containing the correct insert were identified (figure 6.6) and sequenced (section 2.2.10) to check for any PCR induced errors using internal primers (NHSC1Fseq/NHSC1Rseq, table 6.2). No sequence changes were identified in either of the two clones.

Following successful cloning of the C-terminus of NHS, expression of the GST-NHS fusion protein (NHSC1) was induced in JM109 cells (section 2.4.4.1). Total lysates of induced and uninduced cell cultures were resolved by SDS-PAGE (section 2.4.1), transferred to a nitrocellulose membrane (section 2.4.3), and immunoreactive bands detected using the immune, final bleed anti-sera from both rabbits (rabbit 1, SG1703 and rabbit 2, SG1704). Primary antibodies were visualised using a goat anti-rabbit secondary antibody conjugated to horse radish peroxidase (1:30,000, Pierce, UK), and detected using the ECL Plus detection reagent (Amersham Biosciences). The predicted

molecular weight of NHSC1 was 50.1 kDa. No immunoreactive band was observed corresponding to GST-NHSC1 with the immune anti-serum SG1704.

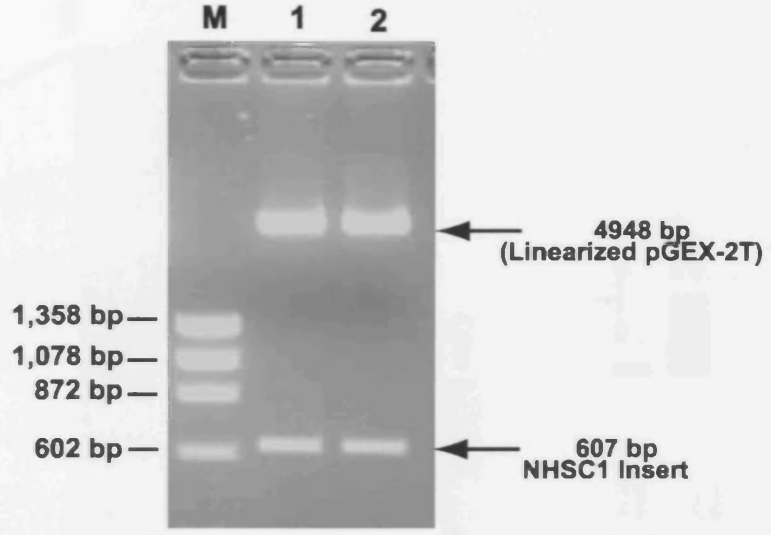


Figure 6.6 Double digest of pGEX-2T clones containing the NHSC1 fragment

The NHSC1 fragment (607 bp) was amplified using primer pair NHSC1F/NHSC1R and cloned into pGEX-2T. Two clones containing the correct insert were identified. M, molecular weight marker; 1, clone 1; 2, clone 2.

In contrast SG1703 detected an immunoreactive band of the predicted molecular weight (approximately 50 kDa, see arrow in figure 6.8) was detected (figure 6.8, B, section 2.4.3). Immunoreactive bands of approximately 50 kDa were detected in both the induced and uninduced cells, although a much stronger immunoreactive band was observed in the induced lanes at approximately 50 kDa (figure 6.8, B). This 50 kDa band was not detected by the pre-immune sera (figure 6.8, A).

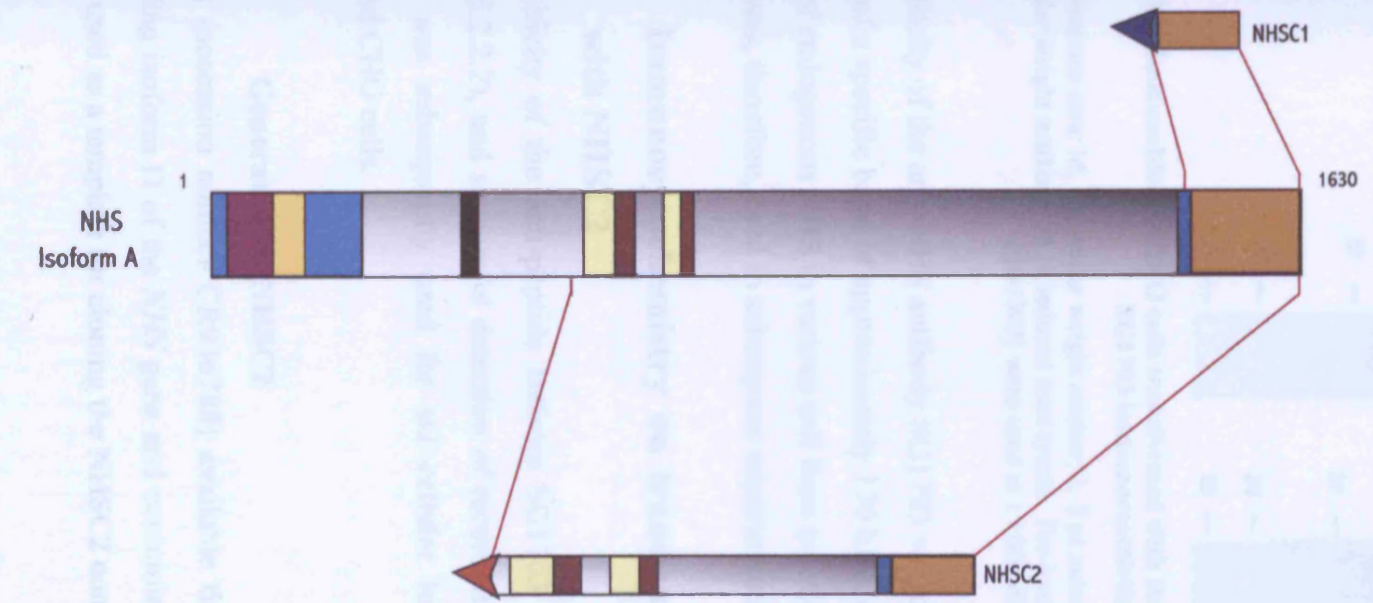


Figure 6.7 NHS constructs generated.

Putative domains in NHS were cloned either into prokaryotic or mammalian expression vectors for subsequent functional analysis. Refer to table 1 in appendix B for more detailed information on NHS constructs generated.

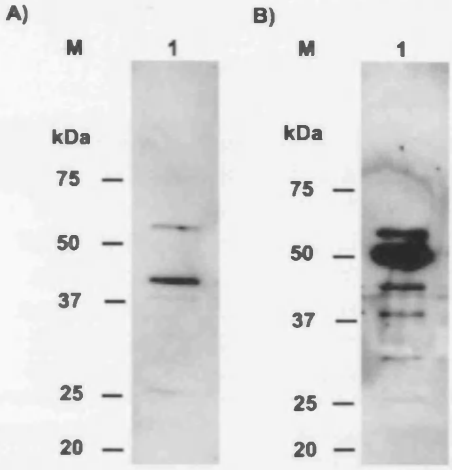


Figure 6.8 Immunoblots of CHO cells transformed with the NHSC1 construct as detected by SG1703 immunoreactivity

A) Pre-immune sera: M, molecular weight marker; 1, 5 µl induced total lysate; B) Immune sera: M, molecular weight marker; 1, 5 µl induced total lysate. Pre-immune and immune sera from rabbit 1 (SG1703) were used at 1:1000 dilution.

The specificity of the anti-NHS antibody SG1703 was determined through peptide blocking of a specific band of approximately 170 kDa, detected through Western blot analysis of endogenous NHS in various cell lines (section 6.2.3.1). Anti-NHS antibody SG1703 was, therefore, used in subsequent experiments.

6.2.2.3 Immunocytochemistry on transiently transfected CHO cells with NHSC2

The specificity of the anti-peptide antisera SG1703 was tested as mentioned above (section 6.2.2.2), and successful detection of recombinant protein on immunoblots, the antibody was subsequently used for sub-cellular localisation studies in transiently transfected CHO cells.

6.2.2.3.1 Generation of NHSC2

A cDNA (accession number CR936788) available through RZPD (Berlin, Germany) representing isoform D of the *NHS* gene and containing exons 1c, 2, 3, and 4-8 (figure 4.9) was used as a template for cloning the NHSC2 construct (figure 6.7).

Primers with engineered restriction endonuclease sites for in frame cloning into pCMV-Tag3b, were designed (section 2.2.2, table 6.3) to PCR amplify (section 2.2.1) the C-terminal half of the *NHS* gene (exons 6-8, amino acids 471-1630, figure 6.7, table 1 in appendix B).

Gene	Primer	Sequence (5'-3')	Annealing temperature (°C)	Amplicon size (bp)
NHS	NHSC2 F	CG**GGATCC**ACCATGTTTACTCCTGCAGTGAGCAG	57	3,496
	NHSC2 R	CG**GAATTC**TGGTACAGGCAGTCTATGTTGAACTC	57	
	NHSC2 Fb	TGAAGCAGAGGGTGTGTTCG	55	
	7Fb	TCGGAGTTGTCACTAAACACAG	55	
	NHSC2 Fc	ATCCTCACTACAGCAACCCTC	55	
	NHSC2 Fd	AAGACCAAGCATCTCTTTCAGG	55	
	NHSC2 Fe	TGCCTGGTACTATCAGCTATG	55	
	NHSC2 Rb	TCCTCAGTTGTTCGAGGTTTG	55	

Table 6.3 Primer sequences used to amplify a 3,496 bp fragment from exon 6-8 of the *NHS* gene. Restriction endonuclease sites engineered into the primer sequences are highlighted in bold and underlined (GGATCC – BamHI, GAATTC – EcoRI).

Using the primers listed in table 6.3, a 3,469 bp fragment from exons 6-8 of the *NHS* gene (coding for the last 1,160 aa of the NHS protein, amino acids 471-1630) was amplified and TA cloned into the pGEM-T Easy vector (section 2.3.4.1). Clones with the correct molecular weight insert were identified through restriction endonuclease digestion of minipreped DNA (section 2.3.3.1), and sequenced (section 2.2.10) using vector sequencing primers (M13 forward and reverse) and internal sequencing primers to check for PCR induced errors.

Following confirmation of the correct sequence for NHSC2, the insert was released from the pGEM-T Easy vector through digestion with restriction endonucleases; PvuI, BamHI, and EcoRI. Since the molecular weight of the pGEM-T Easy vector (3,015 bp) and NHSC2 insert (3,496 bp) were similar, the pGEM-T easy vector was first cut with PvuI. The PvuI restriction endonuclease cut pGEM-T Easy at two sites. This resulted in the release of a 1,099 bp pGEM-T Easy vector fragment, leaving the NHSC2 insert and the remainder of the pGEM-T Easy vector (1,916 bp). Fragments from the digested clone were resolved on a 0.5 % (w/v) agarose gel (section 2.2.5.1) and the larger fragment containing the NHSC2 insert and the remainder of the pGEM-T Easy vector, excised and purified (section 2.2.6). The purified fragment was digested (section 2.2.11) with BamHI and EcoRI (sites engineered into the NHSC2 primer sequences) to release the 3,496 bp NHSC2 insert (figure 6.9). The fragments resulting from this second, double digest were resolved on a 0.5 % (w/v) agarose gel (figure 6.9; section 2.2.5.1) and the NHSC2 insert excised and purified (section 2.2.6).

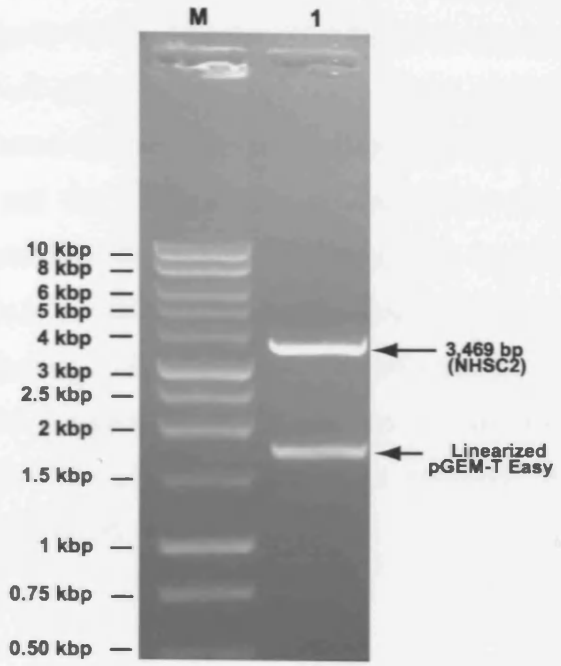


Figure 6.9 Digested pGEM-T Easy/NHSC2 clone, resolved on a 0.5 % (w/v) agarose gel.

M, marker; 1, release of the NHSC2 insert from pGEM-T Easy (pre-digested with PvuI) following digestion with BamHI and EcoRI.

The excised, purified NHSC2 insert was then subcloned in frame (section 2.3), into a mammalian expression vector (pCMV-Tag3b) pre-digested with BamHI and EcoRI. Clones containing the correct insert were verified through digestion with BamHI and EcoRI.

6.2.2.3.2 Expression of the NHSC2 construct in CHO cells

Cloning of the NHSC2 insert into the pCMV-Tag3b vector with N-terminal Myc-tag allowed for expression of an NHSC2 construct (figure 6.7, table 1 in appendix B) and the C-terminal epitope for the anti-NHS antibody SG1703 described in section 6.2.2.1. The pCMV-Tag3b/NHSC2 construct was transiently transfected into CHO cells (section 2.5.5). Transfected cells were then fixed, permeabilised and stained (section 2.5.6) for NHSC2 expression using both SG1703, and a monoclonal antibody to the Myc tag epitope (mAb 9E10, Sigma-Aldrich, UK). Primary antibody binding was visualised using goat anti-rabbit Cy3 and donkey anti-mouse Cy2 for SG1703 and 9E10, respectively. Nuclei of the CHO cells were counterstained with DAPI (section 2.5.6). Images of transfected CHO cells with the Myc-NHSC2 construct were taken using a

Zeiss LSM 510 confocal microscope (figure 6.10). Excitation wavelengths were, 354 nm (DAPI), 488 nm (mAb 9E10), 543 nm (SG1703 Ab).

Expression was confirmed through detection of the Myc tag using mAB 9E10. SG1703 (section 6.2.2), but not the pre-immune antisera (data not shown), also detected expression of the Myc-NHSC2 construct in transiently transfected CHO cells (figure 6.10). Detection of the Myc-NHSC2 construct using SG1703 revealed brighter staining on Myc positive cells only. Results from immunostaining with secondary antibody only, did not result in a signal (data not shown). From these studies, the NHSC2 construct (amino acids 471-1630) appeared to be excluded from the nucleus (nuclei were counterstained with DAPI in figure 6.10).

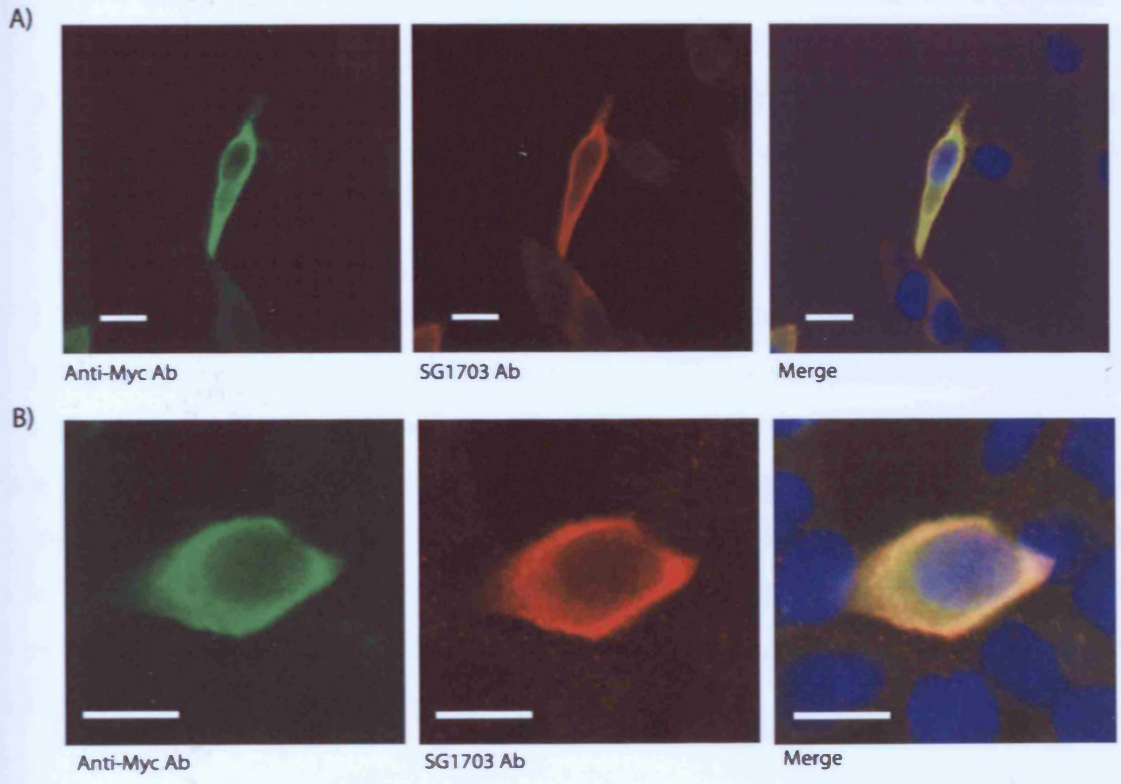


Figure 6.10 Immunofluorescent confocal microscopy of transiently transfected CHO cells with the pCMV-Tag3b/NHSC2 construct.

Anti-NHS antibody SG1703 (1:1000 dilution) detected expression of Myc-NHSC2 in CHO cells. Expression was confirmed by detection of the myc tag with mAb 9E10 (1:1000 dilution). Panel A is of a lower magnification than panel B. Scale bars 10 μm.

6.2.3 Localisation of NHS in metastatic rat mammary adenocarcinoma (MTLn3) epithelial cells

Anti-NHS antibody SG1703 successfully detected NHS construct expression by immunocytochemistry (section 6.2.2.3) and western blot analysis (section 6.2.2.2) the sub cellular localisation of NHS in MTLn3 cells was investigated. MTLn3 cells have been used extensively for studying the actin cytoskeleton, including the localisation of the Arp2/3 complex with an anti-p34 subunit antibody (Bailly *et al.*, 1999). Arp2/3 localises at the leading edge of lamellipods in MTLn3 cells (Bailly *et al.*, 1999, sections 1.9 and 1.10), which can be easily stimulated through the use of epidermal growth factor (EGF).

The homology identified between the NHS family of proteins and the WASP family of proteins (section 6.2.1) has highlighted putative WH2 and CA domains within the NHS protein (figures 6.2, 6.3, and 6.4). The CA domain within the WASP family of proteins facilitates the binding to and activation of the Arp2/3 complex, whilst the WH2 domain binds G-actin bringing actin monomers into close proximity with the Arp2/3 complex. MTLn3 cells provided a useful tool, therefore, for studying the localisation of NHS due to the clarity visualising the actin cytoskeletal structures, and the Arp2/3 localisation studies previously carried out on these cells.

6.2.3.1 Detection of endogenous NHS through immunoblotting

A range of cell lines (CHO, COS-7, Caco-2, MDCK, and MTLn3) were cultured as described in section 2.5.1.1, and grown to confluency. The cells were then lysed and the cell lysates collected. Equal volumes of total cell lysates were then resolved by SDS-PAGE (section 2.4.1.1), blotted (section 2.4.3), and immunoreactive bands detected using anti-NHS antibody SG1703. A strong immunoreactive band of an apparent molecular weight of approximately 170 kDa was detected with SG1703 in all cell lines used in this study (figure 6.11). No immunoreactive band of the same size was detected with the pre-immune serum for each cell line. The predicted molecular weights of the NHS isoforms without any post-translational modifications were: 195.6 kDa (isoform A), 160.2 kDa (isoform B), 176.9 kDa (isoform C), 174.4 kDa (isoform D), and 176.8 kDa (isoform E).

The immunoreactive band detected with SG1703 (figure 6.11, indicated by the arrow) has an apparent molecular weight of approximately 170 kDa. It is unclear which NHS isoform this may represent (A to E) or if it represents a novel alternative isoform. To confirm the specificity of the SG1703 antibody, peptide competition with SG1703 was performed. SG1703 was incubated in the presence of 50 μg/ml of the NHS antigen peptide, for two hours at 4°C with end-on-end rotation (referred to from here as the peptide-blocked immune antiserum). Total cell lysates, from all five cell lines (CHO, COS-7, Caco-2, MDCK, and MTLn3) were resolved by SDS-PAGE (section 2.4.1.1), blotted (section 2.4.3), and immunoreactive bands detected using the peptide-blocked immune antiserum. The immunoreactive band running at approximately 170 kDa detected by the SG1703 antibody in each cell line, was found to be significantly reduced in intensity (figure 6.11, indicated by the arrow) in contrast to other inmmunoreactive bands (figure 6.11, indicated by the asterisks). This data strongly suggested that all five cell lines contain an isoform of endogenous NHS that migrates with an apparent molecular weight of 170 kDa.

6.2.3.2 Immunolocalisation of NHS in MTLn3 cells

The Arp2/3 complex is recruited to the membrane of the extending lamellipod in MTLn3 cells following stimulation with epidermal growth factor (EGF; Bailly *et al.*, 1999). The homology between the NHS family of proteins and the WASP family of proteins (section 6.2.1) suggests that NHS may have a role in binding and activating the Arp2/3 complex. To test this hypothesis stimulated and unstimulated MTLn3 cells were immunostained for endogenous NHS.

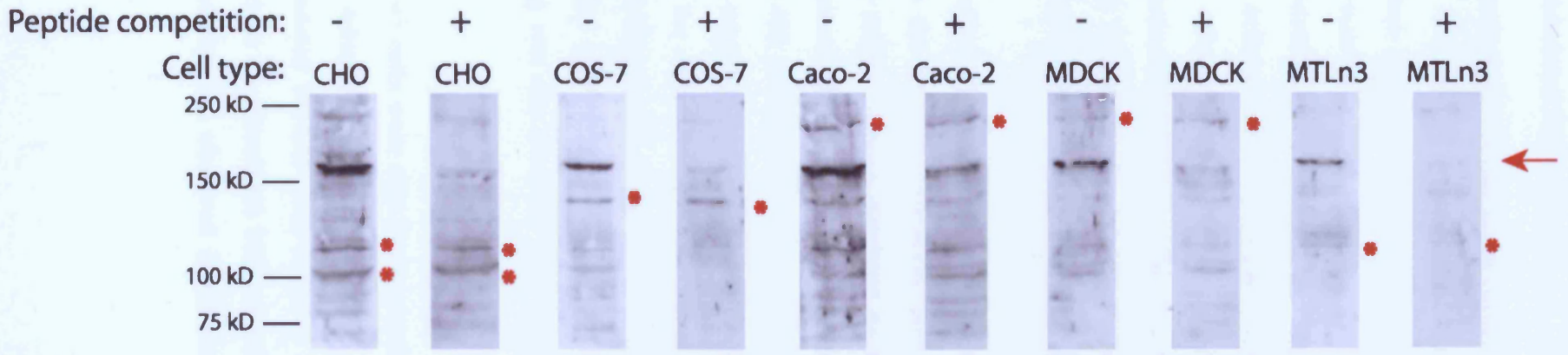


Figure 6.11 Immunoblotting of mammalian cell lysates with anti-NHS antibody SG1703

Red arrow indicates an immunoreactive band corresponding to NHS, migrating at an apparent molecular weight of 170 kDa. Note the intensity of the immunoreactive band corresponding to the NHS protein (indicated by the red arrow) is reduced following peptide competition in contrast to the remaining immunoreactive bands highlighted by red asterisks. CHO, Chinese hamster ovary; COS-7, Monkey African Green Kidney; Caco-2, Human Caucasian colon adenocarcinoma; MDCK, Dog cocker spaniel kidney; MTLn3, Metastatic rat mammary adenocarcinoma.

6.2.3.2.1 Immunofluorescent confocal microscopy of stimulated and unstimulated MTLn3 cells

MTLn3 cells were cultured in MEM alpha containing 5 % (v/v) fetal calf serum, in an atmosphere of 5 % CO_2 at 37°C (section 2.5.1.1). Cells were seeded onto cover slips (section 2.5.7). Once approximately 70 % confluent, MTLn3 cells were starved in starvation media (serum-free MEM alpha/0.35 % (w/v) BSA) for a maximum of 3 hours before being stimulated for 3 minutes in the presence of EGF (section 2.5.7). Stimulated MTLn3 cells were then immediately fixed and stained (section 2.5.6.1) with either the anti-p34 subunit antibody or the anti-NHS antibody SG1703 and the pre-immune serum. Unstimulated MTLn3 cells were fixed and stained with the appropriate antiserum once they were approximately 70 % confluent. The actin cytoskeleton was visualised by staining with Alexa Fluor 488 phalloidin.

6.2.3.2.1.1 *Localisation of endogenous Arp2/3 complex*

Figure 6.12 shows the immunofluorescent staining pattern observed for Arp2/3 in stimulated MTLn3 cells. Arp2/3 localised to the leading edge of the lamellipod, indicated by the arrows (figure 6.12 A and B), as previously reported (Bailly *et al.*, 1999). Alexa fluor 488 phalloidin labelling of the actin cytoskeleton revealed actin rich areas at the leading edge of lamellipods, along with stress fibres (figure 6.12, A and B). The p34 subunit of the Arp2/3 complex therefore, co-localised with actin at the leading edge of stimulated MTLn3 cells (figure 6.12 A and B). There did not appear to be any co-localisation of the p34 subunit with stress fibres (figure 6.12 A and B). Some cytoplasmic staining was also observed with the anti-p34 subunit antibody (figure 6.12 A and B).

Unstimulated MTLn3 cells were also immunostained for endogenous Arp2/3 complex using the anti-p34 subunit antibody (figure 6.12, C). Immunostaining of the Arp2/3 complex in unstimulated MTLn3 cells revealed a greater proportion of cytoplasmic staining in comparison to stimulated MTLn3 cells (figure 6.12, A-C). In addition, strong immunoreactivity was observed at membrane ruffles (indicated by arrows in figure 6.12, C).

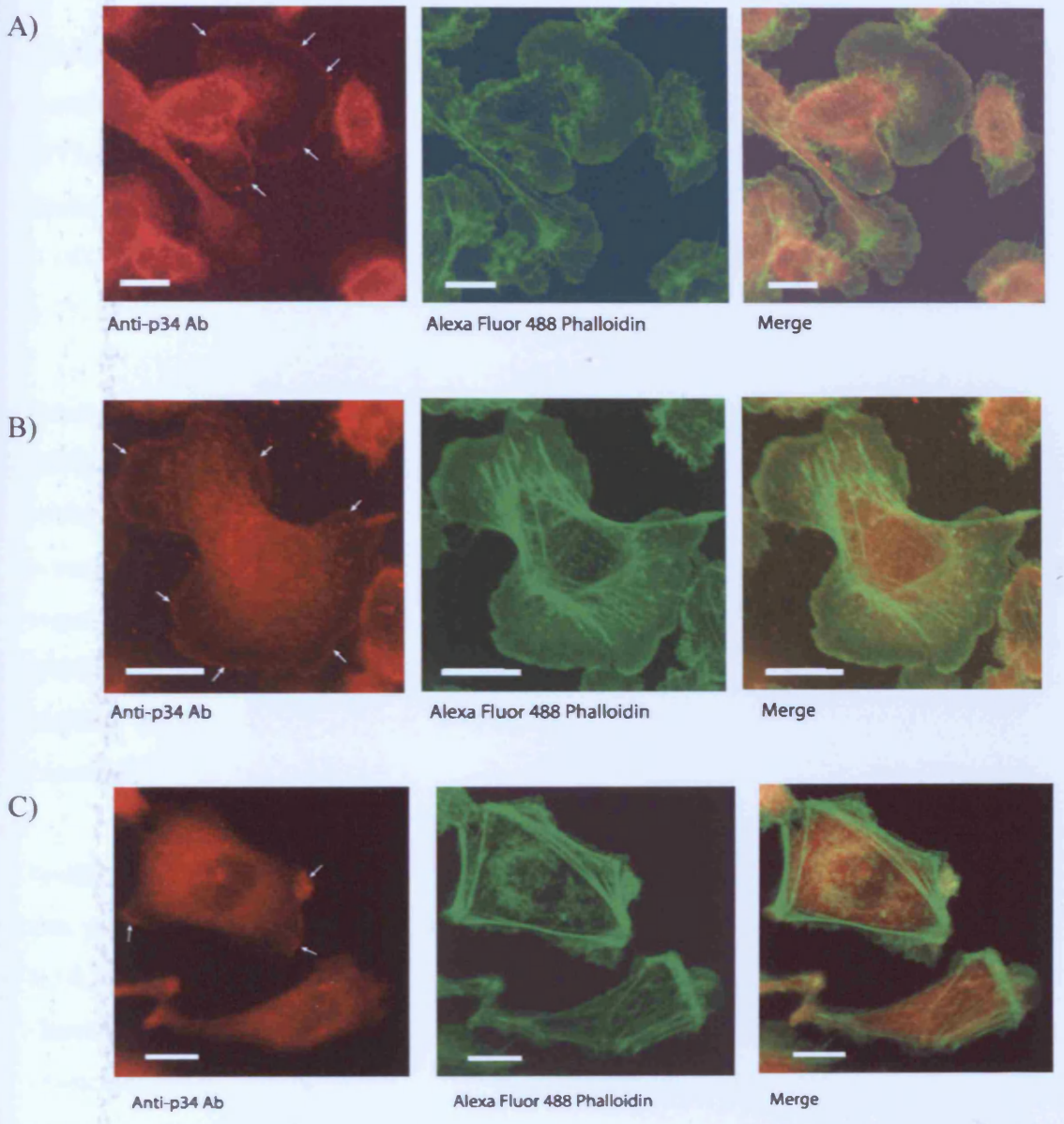


Figure 6.12 Immunofluorescent confocal microscopy of stimulated and unstimulated MTLn3 cells to detect endogenous Arp2/3

Anti-p34 subunit rabbit polyclonal antibody (1:50 dilution), detected expression of endogenous Arp2/3 in MTLn3 cells stimulated with EGF to form lamellipods. Labelling of the actin cytoskeleton with Alexa Fluor 488 phalloidin. Scale bars 10 μm. (A) Endogenous Arp2/3 localised to the leading edge of lamellipods in stimulated MTLn3 cells (arrows) (B) Higher magnification of endogenous Arp2/3 at the leading edge of lamellipods (arrows). (C) Anti-p34 subunit rabbit polyclonal antibody (1:50 dilution) detected expression of endogenous Arp2/3 in unstimulated MTLn3 cells. Arrows indicate membrane ruffles

The cell edge of unstimulated MTLn3 cells was devoid of any staining, consistent with the absence of lamellipodia (figure 6.12, C). Immunostaining with secondary antibody only, for both stimulated and unstimulated MTLn3 cells, had no signal (data not shown).

6.2.3.2.1.2 *Localisation of endogenous NHS*

Localisation of endogenous NHS was investigated in stimulated and unstimulated MTLn3 cells (figures 6.13 and 6.14). SG1703 immunoreactivity was detected in the nucleus of both stimulated and unstimulated MTLn3 cells (figures 6.13, A and B and 6.14). Staining of the nucleus was also observed with the pre-immune serum (figure 6.13, C) however, suggesting that this maybe non-specific.

Interestingly, in stimulated MTLn3 cells (figure 6.13, A and B) but not unstimulated MTLn3 cells (figure 6.14), NHS immunoreactivity was detected at the edges of cells, which had formed a broad flat lamellipod. The staining pattern for NHS closely resembled that of the p34 subunit of the Arp2/3 complex (figure 6.12) at the leading edge of lamellipods (indicated by arrows in figure 6.13, A and B). Thus, endogenous NHS appeared to localise with actin and was similar to Arp2/3 complex staining at the leading edge of lamellipods. No immunofluorescence at the leading edge of lamellipods was observed with the pre-immune serum (figure 6.13, C).

Unlike the localisation pattern observed for the p34 subunit of the Arp2/3 complex in this study (figure 6.12), SG1703 decorated some actin stress fibres (figure 6.13, B and 6.14, B). Therefore, endogenous NHS appeared to be localised along some of the stress fibres (visualised with Alexa Fluor 488 Phalloidin) observed in stimulated MTLn3 cells (indicated by green arrows in figure 6.13, B). Co-localisation of NHS with stress fibres was also observed on occasion in unstimulated MTLn3 cells (indicated by arrows in figure 6.14, B). Decoration of stress fibres was not observed with the pre-immune serum (figure 6.13, C and 6.14, C).

Slightly less cytoplasmic immunoreactivity with SG1703 was observed in stimulated MTLn3 cells (figure 6.13, A and B) when compared to unstimulated MTLn3 cells (figures 6.14). This is analogous to the staining pattern observed for the p34 subunit of the Arp2/3 complex in stimulated and unstimulated MTLn3 cells (figures 6.12).

The cell edge of unstimulated MTLn3 cells was devoid of any staining, consistent with the absence of lamellipodia (figure 6.12, C).

6.2.3.2.1.2 *Localisation of endogenous NHS*

Localisation of endogenous NHS was investigated in stimulated and unstimulated MTLn3 cells (figures 6.13 and 6.14). SG1703 immunoreactivity was detected in the nucleus of both stimulated and unstimulated MTLn3 cells (figures 6.13, A and B and 6.14). Staining of the nucleus was also observed with the pre-immune serum (figure 6.13, C) however, suggesting that this maybe non-specific.

Interestingly, in stimulated MTLn3 cells (figure 6.13, A and B) but not unstimulated MTLn3 cells (figure 6.14), NHS immunoreactivity was detected at the edges of cells, which had formed a broad flat lamellipod. The staining pattern for NHS closely resembled that of the p34 subunit of the Arp2/3 complex (figure 6.12) at the leading edge of lamellipods (indicated by arrows in figure 6.13, A and B). Thus, endogenous NHS appeared to localise with actin and was similar to Arp2/3 complex staining at the leading edge of lamellipods. No immunofluorescence at the leading edge of lamellipods was observed with the pre-immune serum (figure 6.13, C).

Unlike the localisation pattern observed for the p34 subunit of the Arp2/3 complex in this study (figure 6.12), SG1703 decorated some actin stress fibres (figure 6.13, B and 6.14, B). Therefore, endogenous NHS appeared to be localised along some of the stress fibres (visualised with Alexa Fluor 488 Phalloidin) observed in stimulated MTLn3 cells (indicated by green arrows in figure 6.13, B). Co-localisation of NHS with stress fibres was also observed on occasion in unstimulated MTLn3 cells (indicated by arrows in figure 6.14, B). Decoration of stress fibres was not observed with the pre-immune serum (figure 6.13, C).

Slightly less cytoplasmic immunoreactivity with SG1703 was observed in stimulated MTLn3 cells (figure 6.13, A and B) when compared to unstimulated MTLn3 cells (figures 6.14). This is analogous to the staining pattern observed for the p34 subunit of the Arp2/3 complex in stimulated and unstimulated MTLn3 cells (figures 6.12).

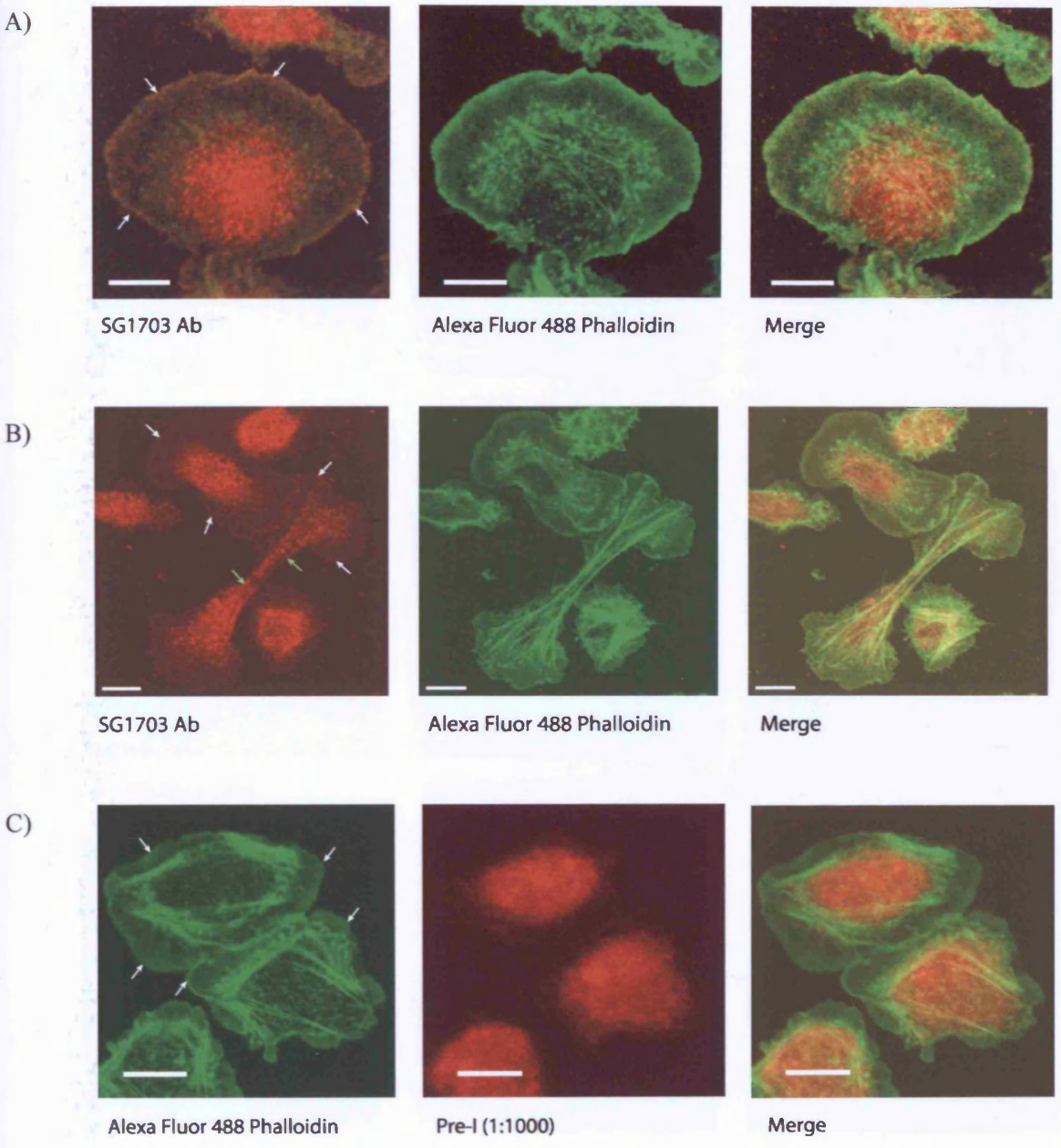


Figure 6.13 Immunoflurescent confocal microscopy of endogenous NHS following stimulation of lamellipod formation in MTLn3 cells

Anti-NHS antibody SG1703 (1:1000 dilution), detected expression of endogenous NHS in MTLn3 cells stimulated with EGF to form lamellipods. Staining of the actin cytoskeleton was achieved using Alexa Fluor 488 phalloidin. (A) Endogenous NHS localised at the leading edge of lamellipods in stimulated MTLn3 cells (B) Lower magnification of localised endogenous NHS to the leading edge of lamellipods in an MTLn3 cell. White arrows indicate localisation of NHS to the leading edge of the lamellipods. Green arrows highlight stress fibres. (C) Pre-immune serum for SG1703 (Pre-I, 1:1000 dilution) does not detect the leading edges of lamellipods. Scale bars 10 μm.

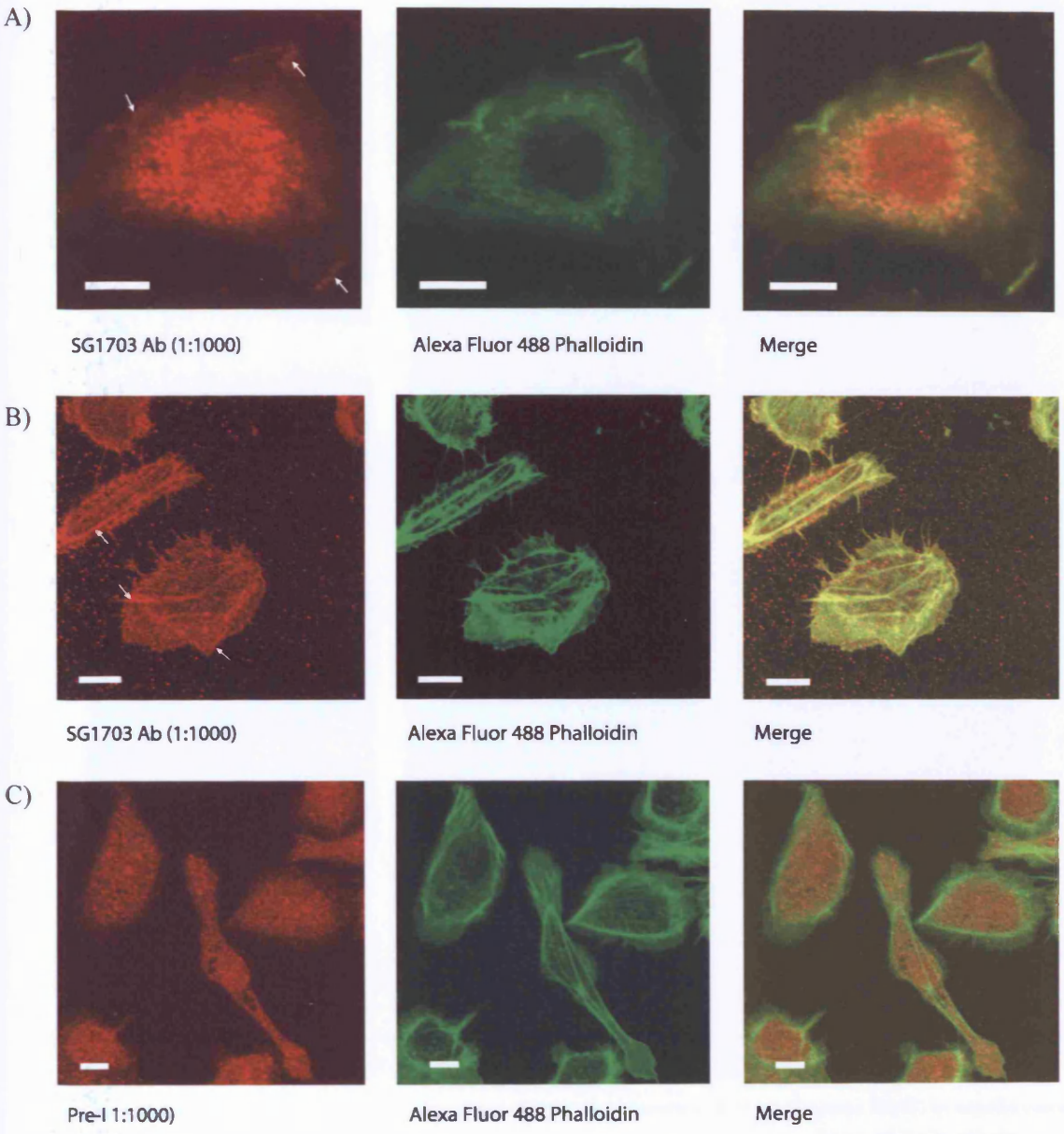


Figure 6.14 Immunofluorescent confocal microscopy showing decoration of some actin fibres and localisation of NHS to membrane ruffles, in unstimulated MTLn3 cells

Anti-NHS antibody SG1703 (1:1000 dilution) detected expression of endogenous NHS in unstimulated MTLn3 cells (A and B). Staining of the actin cytoskeleton was achieved using Alexa Fluor 488 phalloidin. (A) Endogenous NHS localised to membrane ruffles (indicated by the arrows); (B) Endogenous NHS in unstimulated MTLn3 cells, highlighting co-localization with some actin fibres (indicated by the white arrows). (C) Pre-immune serum for SG1703 (Pre-I, 1:1000 dilution) does not detect actin stress fibres. Scale bars 10 μm. The non-specific background spots in B, are a result of the Cy3 labelled secondary antibody.

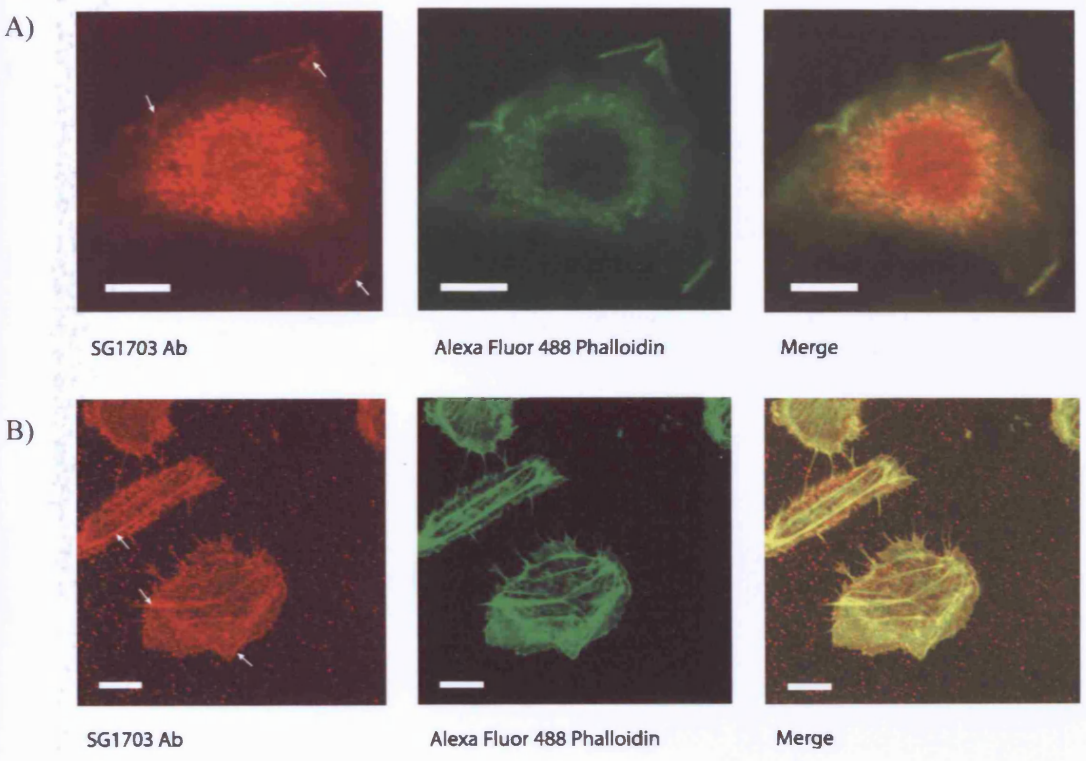


Figure 6.14 Immunofluorescent confocal microscopy showing decoration of some actin fibres and localisation of NHS to membrane ruffles, in unstimulated MTLn3 cells

Anti-NHS antibody SG1703 (1:1000 dilution) detected expression of endogenous NHS in unstimulated MTLn3 cells. Staining of the actin cytoskeleton was achieved using Alexa Fluor 488 phalloidin. (A) Endogenous NHS localised to membrane ruffles (indicated by the arrows); (B) Endogenous NHS in unstimulated MTLn3 cells, highlighting co-localization with some actin fibres (indicated by the white arrows). Scale bars 10 μm. The non-specific background spots in B, are a result of the Cy3 labelled secondary antibody.

Localisation of endogenous NHS to the leading edge of lamellipods in stimulated MTLn3 cells (figure 6.13, A and B) was confirmed by generating a phase contrast image in parallel to imaging localised endogenous NHS in stimulated MTLn3 cells (figure 6.15). The very edges of stimulated MTLn3 cells were clearly defined using phase contrast (figure 6.15), and indeed confirmed endogenous NHS localisation at the leading edges of lamellipoda in stimulated MTLn3 cells.

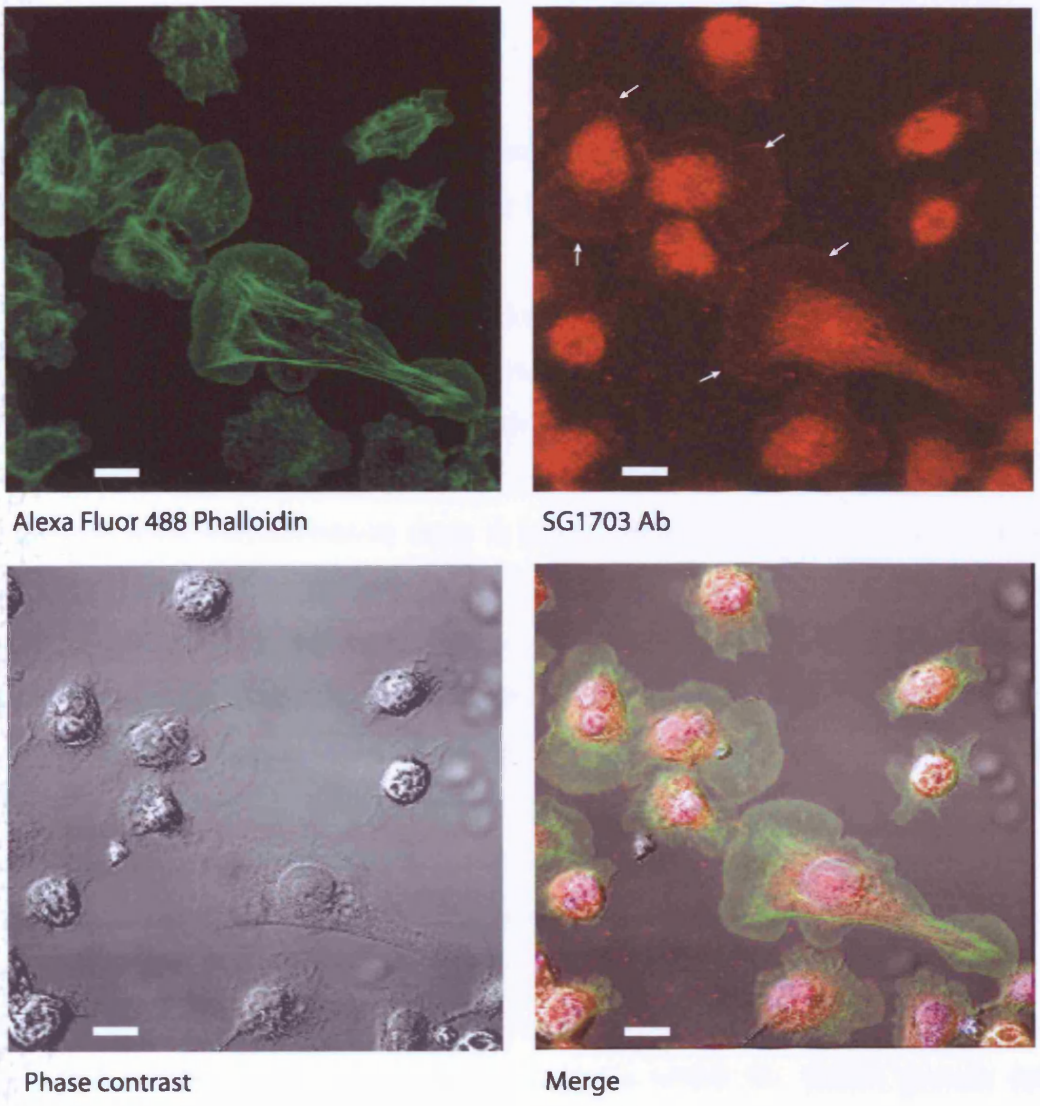


Figure 6.15 Confocal microscopy of endogenous NHS in stimulated MTLn3 cells with defined leading edges

SG1703 (1:1000 dilution) detected expression of endogenous NHS in MTLn3 cells stimulated with EGF to form lamellipods. Staining of the actin cytoskeleton was achieved using Alexa Fluor 488 phalloidin. Phase contrast of the same field defines the edges of stimulated MTLn3 cells. White arrows indicate localisation of NHS to the leading edge of lamellipods. Scale bars 10 μm.

6.2.4 Functional analysis of the putative CA domains in NHS

6.2.4.1 Introduction

Actin polymerisation requires the formation of stable nuclei (the lag phase), followed by rapid addition of actin monomers to the nuclei (the elongation phase; Pollard *et al.*, 2000, section 1.9). The Arp2/3 complex alone has very weak intrinsic actin nucleation ability. However, when activated, the Arp2/3 complex reduces the lag phase by increasing the rate of stable nucleus formation, and increases the rate of the elongation phase. Together, this results in an increase in the rate of actin polymerisation.

The identification of a putative actin binding domain (WH2) in the NHS protein implies that NHS may interact with actin. The pyrenyl actin polymerisation assay is a suitable experiment for testing whether a protein interacts with actin.

The pyrenyl actin polymerisation assay is a very sensitive method for determining if a protein activates the Arp2/3 complex. Actin was conjugated to pyrene through a covalent bond at the cysteine at 374 position (Cytoskelton, US). As a result, the incorporation of pyrene-conjugated actin monomers into an actin filament can enhance fluorescence up to twenty-fold and be detected using a spectrophotometer. Thus, the ability of a protein to activate the Arp2/3 complex (thereby increasing the rate at which actin is polymerised) can be determined by an increase in fluorescence over time. There are numerous examples of nucleation promoting factors that bind and activate the Arp2/3 complex in the literature (e.g. Welch *et al.*, 1998, Machesky et al., 1999; Rohatgi *et al.*, 1999), many of which were determined in part through the use of this actin polymerisation assay. A recent example in which the RickA protein (section 1.12.5) was shown to be a weak activator of the Arp2/3 complex compared with ActA and the WCA domain of N-WASP, is shown in figure 6.16.

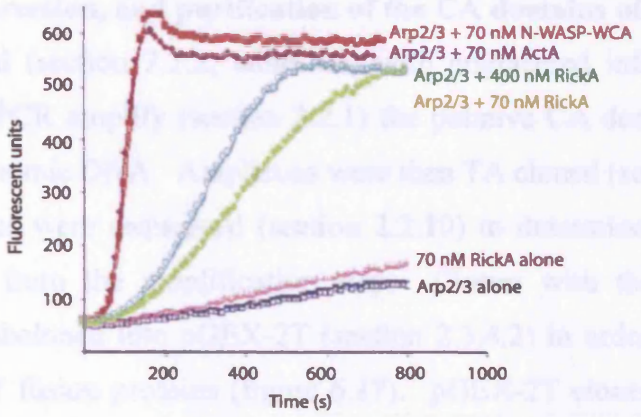


Figure 6.16 Pyrenyl-actin polymerisation assay

An example of activation of the Arp2/3 complex by the WCA domain of N-WASP, ActA, and RickA, resulting in an increased rate of actin polymerisation. Adapted from Gouin *et al.*, 2004.

The WASP protein family members are typical examples of nucleation promoting factors that have been tested for their ability to activate the Arp2/3 complex via the pyrenyl actin polymerisation assay. Indeed the WASP-WCA domain is a commonly used positive control in actin polymerisation assays. Both the WASP-WCA and the N-WASP-WCA are potent activators of the Arp2/3 complex. The V region binds actin monomers, whilst the CA region binds and activates the Arp2/3 complex. Binding of the WCA domain to an actin monomer and the Arp2/3 complex results in the actin monomer being brought into close proximity with the Arp2, and Arp3 subunits, serving as a nucleus for actin filament formation (Kelleher *et al.*, 1995; Machesky and Insall, 1998; Marchand *et al.*, 2001).

The putative WH2 and CA domains of the NHS protein (section 6.2.1) are not contiguous like that of the WASP protein family. In order to test whether the putative CA domains (CA1 and CA2) of the NHS protein are functional, a competitive pyrenyl-actin polymerisation assay was carried out. If the putative CA domains of NHS proved functional, then the presence of such domains in an assay containing the Arp2/3 complex and WASP-WCA domain, would lead to competition between the NHS-CA domain(s) and the WASP-WCA domain for the Arp2/3 complex. Such competition would be predicted to result in a decreased rate of actin polymerisation.

6.2.4.2 Results

6.2.4.2.1 Cloning, expression, and purification of the CA domains of NHS

Primers were designed (section 2.2.2, table 6.4) with engineered inframe restriction endonuclease sites to PCR amplify (section 2.2.1) the putative CA domains (CA1 and CA2) in NHS from genomic DNA. Amplicons were then TA cloned (section 2.3.4) into pGEM-T Easy. Clones were sequenced (section 2.2.10) to determine no PCR errors had been introduced from the amplification step. Clones with the correct insert sequence were then subcloned into pGEX-2T (section 2.3.4.2) in order to generate N-terminally tagged GST fusion proteins (figure 6.17). pGEX-2T clones containing the correct insert were confirmed through double digests (section 2.2.11) of miniprep DNA (section 2.3.3.1) resolved on a 0.5 % (w/v) agarose check gels (section 2.2.5). Clones with the correct insert and no sequence errors were then selected for subsequent expression of the GST fusion protein (section 2.4.4).

Gene	Primer	Sequence (5'-3')	Annealing temperature (°C)	Amplicon size (bp)
NHS	NHS-CA1 F	CG**GGATCC**AGCAGCTCCACGTGCCC	59	357
	NHS-CA1 R	CG**GAATTC**GTAATTCCACTCACTGTCAC	59	
	NHS-CA2 F	CG**GGATCC**GTTGGCGCTAAACCCTCAG	59	266
	NHS-CA2 R	CG**GAATTC**CCCACTCCAGTTGCCACT	59	

Table 6.4 Primer sequences used to amplify the putative NHS CA1 and CA2 domains. Restriction endonuclease sites engineered into the primer sequences are highlighted in bold and underlined (GGATCC – BamHI; GAATTC – EcoRI).

JM109 competent cells were transformed with the NHS-CA1 and NHS-CA2 constructs (section 2.3.2). Transformed cells were then induced to express the GST-fusion proteins (section 2.4.4). Total cell lysates, supernatant, and pellet fractions were then resolved by SDS-PAGE (section 2.4.1.1) and the protein bands visualised with coomassie brilliant blue (section 2.4.2) in order to determine the expression and relative solubility of the GST fusion proteins (figures 6.18, A). The molecular weight of the two fusion proteins were predicted to be 36.7 kDa and 39.9 kDa for NHS-CA1 and NHS-CA2, respectively (see appendix B, table 1). Both the NHS CA1 and CA2 fusion proteins were found to be soluble with little or no protein detected in the pellet fraction. The GST fusion proteins were then purified by glutathione sepharose

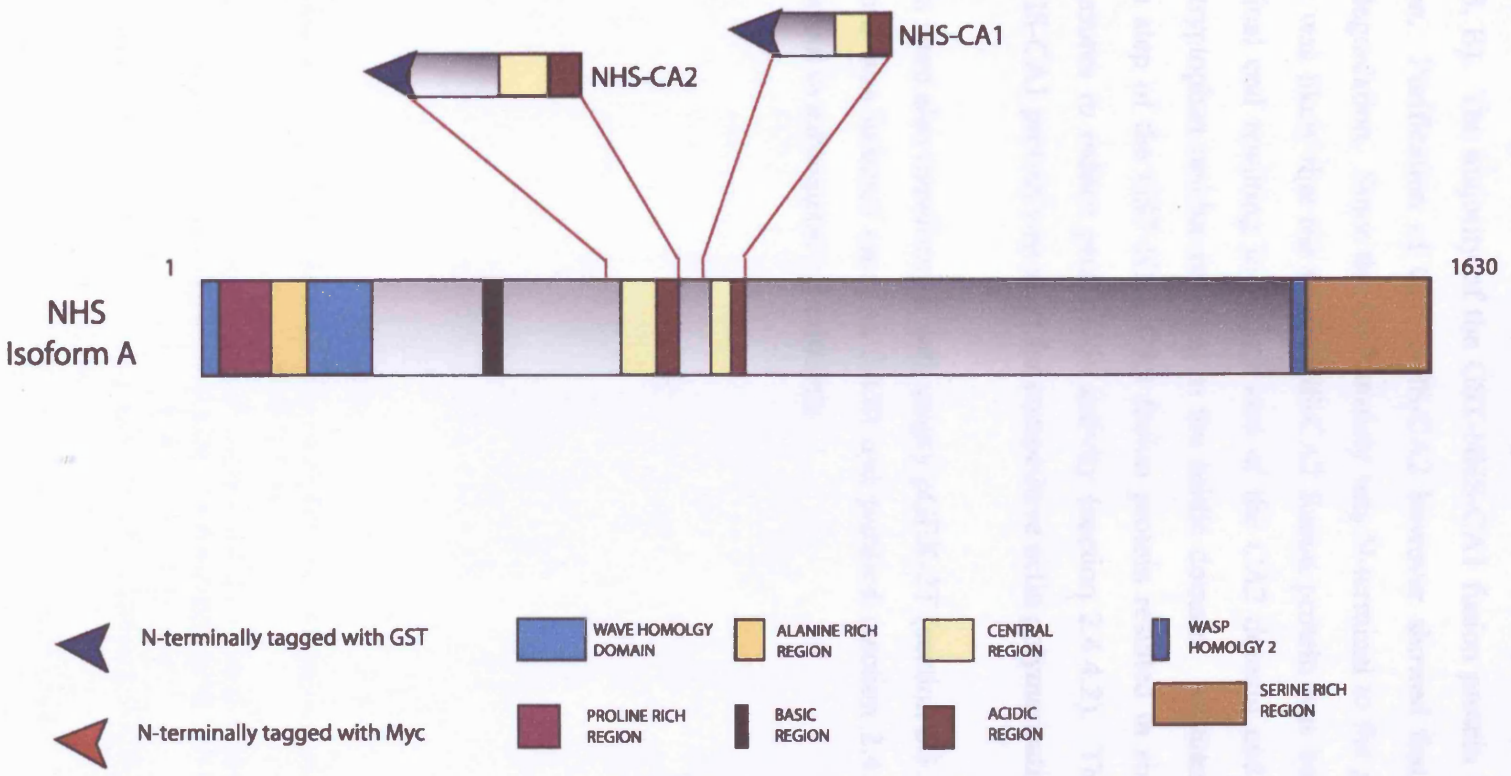


Figure 6.17 NHS constructs generated throughout this research project.

Putative domains in NHS were cloned either into prokaryotic vectors for subsequent functional analysis. Refer to table 1 in appendix B for more detailed information on the NHS constructs generated

affinity chromotography (section 2.4.4.2) for use in pyrenyl actin polymerisation assays (figures 6.18, B).

GST-NHS-CA1 was purified with little degradation during the purification process (figure 6.18, B). The majority of the GST-NHS-CA1 fusion protein was eluted in the first fraction. Purification of GST-NHS-CA2 however showed that the protein was subject to degradation. Since the GST moiety was N-terminal to the putative NHS CA domains, it was likely that the GST-NHS-CA2 fusion protein was being cleaved from the C-terminal end resulting in partial loss of the CA2 domain and in particular the conserved tryptophan residue residing in the acidic domain. An attempt to repeat the purification step of the GST-NHS-CA2 fusion protein resulted in similar degradation despite measures to reduce proteolytic activity (section 2.4.4.2). Therefore, only the purified NHS-CA1 protein was used for competitive actin polymerisation assays.

JM109 cells were also transformed with empty pGEX-2T (section 2.3.2) and expression of GST alone was induced (section 2.4.4) and purified (section 2.4.4.2) for use as a negative control in subsequent experiments.

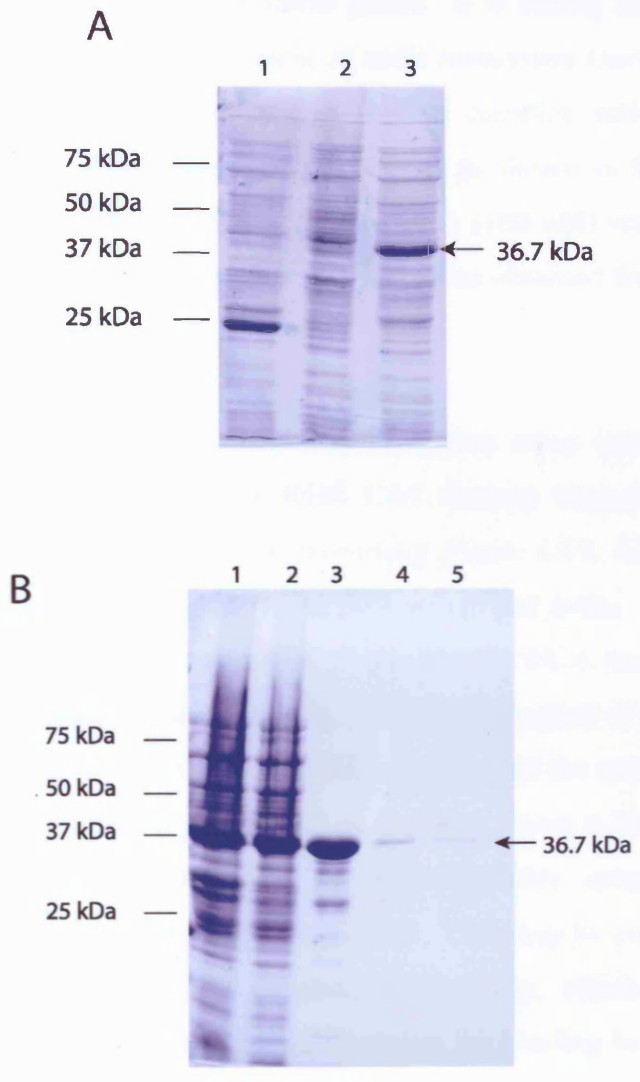


Figure 6.18 Coomassie brilliant blue® stained SDS PAGE gels showing expression and purification of NHS CA1.

(A) Induced GST-NHS CA1 expression: Lane 1) empty pGEX-2T expressing GST, induced total cell lysate 2) Empty pGEX-2T, uninduced total cell lysate; 3) Induced GST-NHS CA1 expression (total cell lysate).

(B) Purification of GST-NHS-CA1. 1) Total cell lysate, 2) Supernatant, 3-5) Eluted fractions 1-3.

6.2.4.2.2 Pyrenyl-actin polymerisation assay

The pyrenyl-actin polymerisation assay was carried out as described in section 2.6. Actin assembly kinetics exhibit an initial lag phase corresponding to the kinetic barrier to nucleation, followed by a rapid linear phase. It is during the latter stage that actin filaments elongate through the addition of actin monomers (section 1.9). When in the presence of the Arp2/3 complex and an Arp2/3 complex activator, the lag phase is reduced and the rate of polymerisation increased, as shown in figure 6.19, A (compare the actin only curve with that of the WASP-WCA (100 nM) with actin and the Arp2/3 complex). A GST-WASP-WCA fusion protein was obtained from Cytoskelton Inc. for use as a positive control.

Initial results from the pyrenyl-actin polymerisation assay (section 2.6) in which the WASP-WCA domain and putative NHS CA1 domain were allowed to compete for binding to the Arp2/3 complex, were promising (figure 6.19, A). This experiment was conducted in the presence of 1 % (w/w) pyrene-labelled actin. The initial lag phase in the competitive assay compared to that of the WASP-WCA was extended and the rate of the elongation phase reduced slightly. The total amount of polymerised actin over the same time period between the competitive assay and the non-competitive assay was found to differ (compare the plateaus of each curve, figure 6.19, A). The total amount of polymerised actin was reduced in the competitive assay despite the original concentration of actin in each assay being equal. This may be explained by the presence of the WASP-WCA domain in the competitive assay, which is unable to bind the Arp2/3 complex due to the potential competition for binding by GST-NHS-CA1. Free WASP-WCA may be removing the total amount of free actin monomers from the assay through its actin-binding domain. Therefore, the rate of elongation was reduced due to the reduction in available Arp2/3 complex, which was bound to the NHS-CA1 domain, and the total amount of polymerised actin reduced due to the removal of free actin monomers from the reaction. The same affect would presumably be observed upon decreasing the concentration of Arp2/3 complex in the assay. Since the NHS-CA1 protein does not contain an actin-binding domain, any

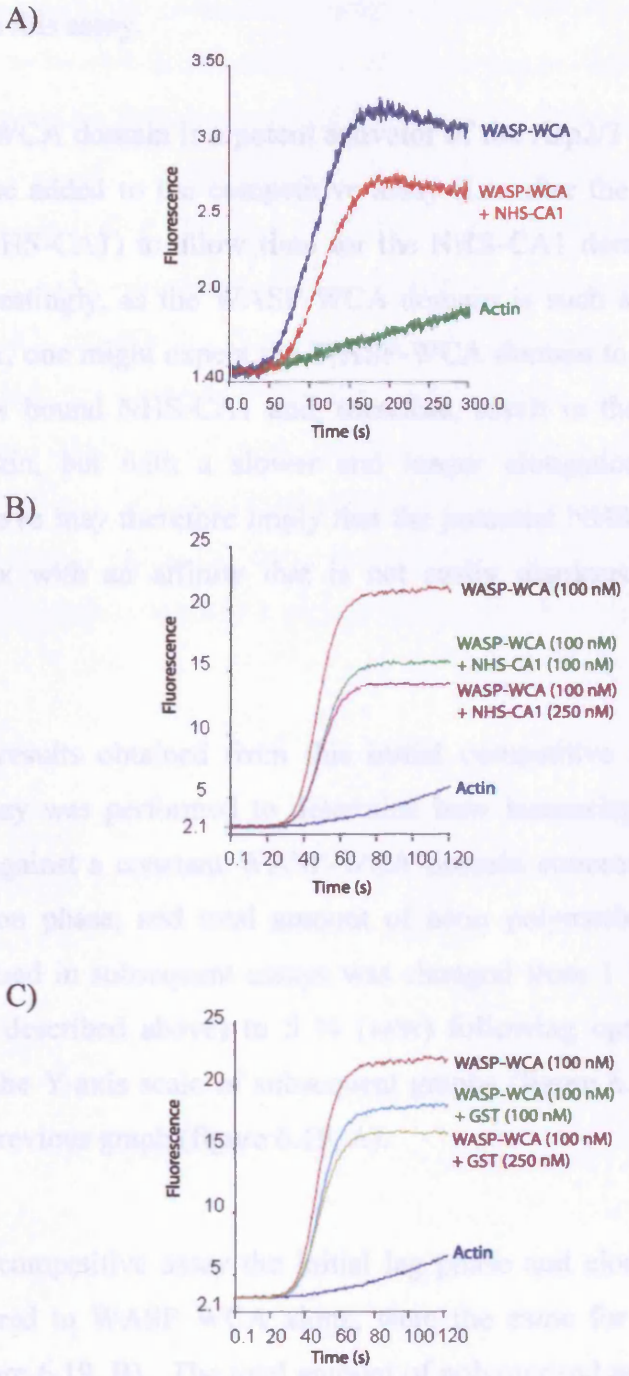


Figure 6.19 Pyrenyl-actin polymerisation assays.

Labelled pyrene actin with either:

(A) Actin alone (1 % (w/w) labelled, blue), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM, blue), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM) + NHS CA1 (100 nM, red)

(B) Actin alone (5 % (w/w) labelled, blue), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM, brown), or Arp2/3 complex (13 nM) + WASP-WCA (100nM) + NHS CA (100 nM, green), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM) + NHS CA (250 nM, purple).

(C) Actin alone (5 % (w/w) labelled, blue), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM, brown), or Arp2/3 complex (13 nM) + WASP-WCA (100nM) + GST (100 nM, light blue), or Arp2/3 complex (13 nM) + WASP-WCA (100 nM) + NHS CA1 (250 nM, mustard yellow).

Arp2/3 complex bound to the NHS-CA1 domain would be unavailable for binding WASP-WCA in this assay.

As the WASP-WCA domain is a potent activator of the Arp2/3 complex, it was the last component to be added to the competitive assay (i.e. after the addition of the Arp2/3 complex and NHS-CA1) to allow time for the NHS-CA1 domain to bind the Arp2/3 complex. Interestingly, as the WASP-WCA domain is such a potent activator of the Arp2/3 complex, one might expect the WASP-WCA domain to eventually displace any Arp2/3 complex bound NHS-CA1 and, therefore, result in the same total amount of polymerised actin, but with a slower and longer elongation step. The resulting experimental curve may therefore imply that the potential NHS-CA1 domain binds the Arp2/3 complex with an affinity that is not easily displaced by the WASP WCA domain.

Following the results obtained from this initial competitive assay, a dose response competitive assay was performed to determine how increasing amounts of the NHS-CA1 domain, against a constant WASP-WCA domain concentration, affected the lag phase, elongation phase, and total amount of actin polymerisation. The amount of labelled actin used in subsequent assays was changed from 1 % (w/w, as used in the previous assay described above) to 5 % (w/w) following optimisation of the assay. Consequently, the Y-axis scale of subsequent graphs (figure 6.19, B and C) differs to the that of the previous graph (figure 6.19, A).

In this second competitive assay the initial lag phase and elongation phase, although reduced compared to WASP WCA alone, were the same for both concentrations of NHS-CA1 (figure 6.19, B). The total amount of polymerised actin was further reduced with increasing concentration of NHS-CA1 (in this particular case, a 2.5 fold increase in NHS-CA1 concentration, figure 6.19, B). This implied that less Arp2/3 complex was available for the WASP-WCA domain to bind at the higher concentration of NHS-CA1 due to NHS-CA1 sequestering the Arp2/3 complex.

To ensure that the observed affect was due to the NHS-CA1 domain and not an experimental artefact, a second assay was designed in which GST alone was placed in

the assay with the WASP-WCA domain and the Arp2/3 complex (figure 6.19, C). The same affect observed in the competitive assay of WASP-WCA and NHS-CA1, was seen in this assay with WASP-WCA and GST alone (compare figures 6.19, B and C). It would appear, therefore, that the observed affect was a result of the GST moiety and not the NHS-CA1 domain. In particular, GST fusion proteins have the potential to dimerise, which may affect the actin assay as the WASP-WCA domain was also GST bound. However, GTS tagged-proteins have routinely been used in these assays and in one case the GST moiety was found to enhance the ability of the WAVE1-WCA domain to activate the Arp2/3 complex compared with His-tagged WAVE1-WCA (Yamaguchi *et al.*, 2002). Another explanation maybe that a contaminant present in the purified NHS-CA1 and GST purified fractions had an effect on this very sensitive actin polymerisation assay. Overall, determining whether the NHS-CA1 domain was functional by competitive assays will require further development.

6.2.5 Discussion

Little was known about the potential function of the NHS protein following its identification in Nance-Horan syndrome. Burdon and colleagues (2003) identified four putative monopartite nuclear localisation signals (NLS) at amino acid positions: 371-379 (RRRKLRRRK), 438-444 (PSRRRIR), 822-825 (RKPK), 1026-1034 (PGGSKRKPK), but reported no significant homology to any known protein or class of protein (Burdon *et al.*, 2003).

Despite there being no report of a potential function for the NHS protein (Burdon *et al.*, 2003), Katoh and Katoh (2004) published a report detailing homology identified between the NHS and NHSL1 proteins, and the Drosophila Gukh (guanylate-kinase holder) protein. Katoh and Katoh sought human homologs of the *gukh* gene, which encodes an adaptor protein bringing together the Drosophila proteins Dlg (Disc large) and Scrib (Scribble). Dlg and Scrib are implicated in the establishment and maintenance of epithelial polarity (Johnson and Wodarz, 2003; Tanentzapf and Tepass, 2003; Humbert *et al.*, 2003). The homology between Drosophila Gukh and the NHS family members, NHS and NHSL1, is quite low. Five regions of homology were identified, of which three reside within the homology domains 2, 6, 7 and 9 (figures 5.5 and 5.6) of the NHS protein family. Whether the homology identified implies that NHS and NHSL1 share some domains with the Gukh protein will need to be determined.

Through bioinformatic analysis of the protein sequences of the NHS family described in this thesis, in particular NHS, NHSL1, and NHSL2, homology to the WASP protein family was identified.

The role of the WASP protein family in regulating the actin cytoskeleton is currently a vibrant area of research, especially since WASP was shown to activate the Arp2/3 complex (Machesky and Insall, 1998). The WASP protein family form two subgroups: WASP and N-WASP, and WAVE1-3.

The NHS protein family (except NHSL3) share homology at the N-terminus with the WHD of the WAVE proteins. At the time of identifying a potential WHD in the NHS protein family, little was known about WHD function in the WAVE proteins. Recently,

two reports have been published describing the WHD as a binding domain for the Abi protein family (Echarri *et al.*, 2004; Leng *et al.*, 2005). The critical residues for mediating binding to the Abi protein family were mapped within the WHD, through a series of binding experiments in which WAVE1 deletion mutants were used (Echarri *et al.*, 2004; Innocenti *et al.*, 2004). Abi1 has been implicated in the translocation of WAVE1 and WAVE2 to the leading edge of motile cells (Echarri *et al.*, 2004; Innocenti *et al.*, 2004), through binding the WHD in the WAVE proteins. Furthermore, the Arp2/3 complex is localised to the leading edge of motile cells (Bailly *et al.*, 1999), thus Abi1 may translocate the WAVE1 and WAVE2 proteins to the leading edge of motile cells bringing these nucleation promoting factors into the vicinity of the Arp2/3 complex, resulting in localised actin polymerisation at the leading edge.

The Abi-binding domain within the WHD of the WAVEs contains a number of residues conserved across species. These residues are either conserved across the NHS protein family or are substituted by amino acids with similar chemical properties (figure 6.1). The putative WHD in the NHS protein is also highly conserved across species (figure 5.1). Thus, from preliminary analysis, it would appear that NHS, NHSL1, and NHSL2 may contain an Abi-binding domain. Whether or not this potential Abi-binding domain in the NHS protein family is functional remains to be determined. To date, no WHD has been identified in NHSL3. NHSL3 appears to be the most distantly related.

In addition to a putative WHD in the NHS protein family, proline rich regions were found to reside in all members of the NHS protein family except NHSL2. The length of the proline rich regions in the NHS protein family varies considerably from a stretch of 559 amino acids in NHSL3 at its C-terminus (figure 6.2 and table 6.1), to 65 amino acids at the N-terminus of NHS (figure 6.2 and table 6.1). The proline rich region of NHS resides within the potential WHD, just prior to the putative Abi-binding domain. NHSL1 contains a proline rich region towards its C-terminus, spanning 144 amino acids. Proline rich domains are present in all members of the WASP protein family and are reported to bind SH3 domain containing proteins such as Grb2/Ash, Nck, and WISH (Carlier *et al.*, 2000; Fukouka *et al.*, 2001; Rohatgi *et al.*, 2001). Additional SH3 domain containing proteins such as Grb2/Ash, Nck, and WISH, are known to regulate the activity of N-WASP through its proline rich domain (Carlier *et al.*, 2000; Fukouka *et al.*, 2001; Rohatgi *et al.*, 2001). In light of this, the proline-rich domains of the NHS

protein family may bind SH3 domain containing proteins, which would regulate their function within the cell. Since the WAVE proteins bind profilin and the NHS protein family share sequence similarity with the WAVE proteins, it will be important to determine whether any of the NHS protein family members bind the SH3 domain containing protein profilin. Profilin plays an essential role in maintaining a reserve pool of unpolymerised actin for subsequent incorporation into existing filaments (Goldschmidt-Clermont *et al.*, 1991).

Following the identification of a putative WHD and proline-rich regions in the NHS protein family, other WASP protein family domains were sought within the protein sequences of the NHS protein family. As a result, potential CA domains were identified in all four members of the NHS protein family (figures 6.2 and 6.3). Two potential CA domains (CA1 and CA2) were identified in the NHS protein (figures 6.2 and 6.3). CA1 aligned without interruption with the CA domain of N-WASP. The key tryptophan residue within the acidic domain, which is crucial for the binding of the WASP protein family to the Arp2/3 complex, was found to be conserved in all the putative CA domains of the NHS protein family. In addition, arginine and leucine residues within the central domain of the WASP protein family, which facilitate activation of the Arp2/3 complex, were conserved in the putative CA domains of the NHS protein family (figure 6.3).

A putative G-actin binding (WH2) domain was also identified in three members of the NHS protein family (NHS, NHSL1, and NHL2, figure 6.4). The potential WH2 domain in NHS, NHSL1, and NHSL2 resides within a highly conserved domain of these three NHS protein family members (homology domain 8 in figures 5.5 and 5.6). NHSL3 does not share this highly conserved domain.

Unlike the WASP protein family, the putative WH2 domain in the NHS protein family is not positioned next to the putative CA domains (except for NHSL1, figure 6.2). The domain organisation of NHSL1 more closely resembles that of the WAVE proteins, and interestingly the putative Abi-binding domain of NHSL1 shares the greatest sequence identity with the WAVE proteins Abi-binding domain (figure 6.1). The putative CA and WH2 domains of NHSL1 reside close to the C-terminus, as is the case in the WASP protein family. The putative WH2 domains of NHS and NHSL2 also reside at the C-

termini of these proteins, but the putative CA domains are found closer towards the N-terminus. It would appear, therefore, that the NHS protein family have undergone considerable domain shuffling. Importantly, the NHS proteins are much larger than the WASP proteins and therefore, one can envisage that protein folding of the NHS proteins could bring the relevant domains in close proximity.

From the known activators (table 1.5) of the Arp2/3 complex, it has become increasingly apparent that not all nucleation promoting factors have the same central and acidic domain organisation. For example, the bacterial Arp2/3 activator ActA from Listeria *monocytogenes*, contains a central domain downstream of the acidic domain (figure 6.3). What is more, the level of conservation between central and acidic domains of known activators of the Arp2/3 complex varies considerably (figure 6.3). In addition, not all nucleation promoting factors contain a G-actin binding domain but instead have domains that bind F-actin, such as cortactin and Abp1p (section 1.12.3 and 1.12.4). Thus, the potential for an F-actin binding domain in NHSL3 (for which no WH2 domain could be identified) should still be investigated.

As in the WASP protein family, the NHS protein family were found to contain a basic motif residing towards the N-terminus of NHS, NHSL1, and NHSL3, and more centrally in NHSL2. The basic region in WASP and N-WASP binds PIP2 enhancing their ability to activate the Arp2/3 complex (Higgs and Pollard, 2000; Rohatgi *et al.*, 2000), whilst the basic region in WAVE2 was recently reported to bind PtdIns(3,4,5)P_3, which is essential for lamellipod formation (Oikawa *et al.*, 2004). Whether the basic regions in the NHS protein family are able to bind either PIP2, PtdIns(3,4,5)P_3, or another phosphoinositide remains to be determined. Localisation of the NHS protein to the leading edges of lamellipods in stimulated MTLn3 cells strongly supports this idea.

Besides the basic domain and the putative WHD, CA, WH2, domains in the NHS protein family, all four members were found to contain a serine-rich domain, which is not present in the WASP protein family (figure 6.2). The serine-rich region, like the proline-rich region in the NHS protein family, varies in size and location. NHS contains a serine-rich domain spanning 189 amino acids at the very C-terminus, whereas NHSL2 contains a serine-rich domain spanning 24 amino acids and is located more centrally (table 6.1 and figure 6.2). A fourth domain found only to be present in

the NHS protein, was an alanine-rich region residing within the WHD, after the proline-rich domain and overlapping with the first 13 amino acids of the putative Abi-binding domain.

In order to determine whether or not the putative CA domains in the NHS protein are functional and therefore bind and activate the Arp2/3 complex, a pyrenyl-actin polymerisation assay was tested. To test the potential function of the NHS CA domains a competitive assay was tested. The GST-NHS-CA2 fusion protein was degraded and consequently was not used in the competitive assays. Instead, only the ability of the CA1 domain to compete with the WASP-WCA domain for binding the Arp2/3 complex was tested. Initial results implied that the CA1 domain of NHS could compete with the WASP-WCA domain for binding the Arp2/3 complex (figure 6.19, A and B). However, upon testing the effect of GST alone in the assay with the WASP-WCA domain, a similar result was observed (figure 6.19, C).

A possible way to address the problem observed here, would be to cleave the GST moiety from the NHS-CA1 domain using thrombin prior to its use in the assay. Alternatively, the portion of the NHS protein containing the entire putative WH2 and CA domains should be cloned into pGEX-2T for subsequent testing in a non-competitive actin polymerisation assay. If a significant increase in fluorescence over time was observed for pyrene-conjugated actin when in the presence of the Arp2/3 complex and a larger NHS construct containing the putative WH2 and CA domains, one could propose that the WH2 and one (or both) of the CA domains are functional. Since the WASP-WCA domain was obtained from Cytoskeleton Inc., the purification procedure may have been different to that used here to purify the NHS-CA1 protein. For a more comparable assay, it may be worth expressing and purifying a GST-WASP-WCA protein 'in house' so as to use the same purification procedures for all fusion proteins tested.

The actin polymerisation assays could then be complemented with a competitive assay in which the GST moiety is cleaved prior to introducing the CA domains of the NHS protein into the assay. Together it may help to clarify which, if any, of the putative CA domains are capable of binding the Arp2/3 complex. GST pull down assays could be

performed to determine whether NHS directly binds the Arp2/3 complex and globular actin.

Once the domains have been identified and characterised, site directed mutagenesis could be carried out to generate mutant constructs in which key amino acids (such as the conserved tryptophan, figure 6.3) are replaced. If the loss of the conserved tryptophan correlates with a loss in ability to bind and/or activate the Arp2/3 complex, then this would provide more evidence towards a role for NHS in activating the Arp2/3 complex and regulating actin dynamics. Furthermore, if an NHS construct containing the putative WH2, CA1 and CA2 domains is able to activate the Arp2/3 complex, site directed mutagenesis will enable the determination of which of the two putative CA domains are functional.

An antibody was raised against the C-terminal peptide derived from human NHS protein (figure 6.5). The C-terminus of NHS is highly conserved across species (figure 5.1) and, therefore, the anti-peptide antibody should potentially cross react with other species. The high conservation at the C-terminus of NHS across species however, had the potential to generate problems with trying to raise a C-terminal antibody against the NHS protein. It was not known, whether an immune response would be triggered upon exposure to the C-terminal peptide of the human NHS protein sequence. Fortunately, of the two rabbits immunised with the NHS C-terminal peptide, one (SG1703) produced an immune response and the antiserum was used for subsequent immunoblotting and immunocytochemistry experiments in this study.

The last 200 amino acids of the NHS protein fused with a GST tag (NHSC1; figure 6.7) was generated to characterise the NHS C-terminal anti-peptide antibody. The anti-NHS antibody SG1703 detected an immunoreactive band resolving at approximately 50 kDa, as anticipated, on immunoblots (figure 6.8, B). The control pre-immune serum did not detect this species (figure 6.8, A).

An N-terminally Myc-tagged NHS construct (NHSC2, amino acids 471 to 1630; figure 6.7) containing the C-terminal epitope for the anti-NHS antibody SG1703, was successfully generated for transfection into CHO cells. Counterstaining with the mAb 9E10, which detects the Myc tag at the N-terminus of the NHSC2 fusion protein, confirmed the localisation pattern observed with SG1703 (figure 6.10). SG1703

detected expression of NHSC2 in transiently transfected CHO cells (figure 6.10). Expression of NHSC2 appeared to be diffuse cytoplasmic and was excluded from the nucleus. These results demonstrate that SG1703 can successfully detect expression of the NHS protein through immunoblotting and immunocytochemistry.

Interestingly, the NHSC2 construct contains two of the four putative monopartite nuclear localisation signals (822-825 (RKPK), 1026-1034 (PGGSKRKPK)) identified in the NHS protein sequence (Burdon *et al.*, 2003). The absence of the NHSC2 fragment from the nucleus implies that the two putative monopartite nuclear localisation signals residing at amino acid positions: 822-825 and 1026-1034, may not be functional or could be regulated by another stimuli.

To begin to determine whether the homology between the NHS protein family and the WASP protein family was significant, the localisation of endogenous NHS in MTLn3 cells was investigated. MTLn3 cells have been used extensively to study actin cytoskeletal structures and the Arp2/3 complex.

NHS was detected in MTLn3 cells through immunoblotting (figure 6.11). An immunoreactive band of approximately 170 kDa was detected by SG1703 and was not detected with the pre-immune serum (figure 6.11). Peptide competition reduced the intensity of this immunoreactive band, whilst other non-specific bands remained unaffected (figure 6.11). This confirmed the presence of endogenous NHS in MTLn3 cells, and validated their use for the localisation of endogenous NHS. In addition, other cell lines were examined for NHS eexpression. Strong immunoreactive bands of a similar size to that detected in MTLn3 cells (approximately 170 kDa) were detected with SG1703 in Caco-2, CHO, COS-7, and MDCK cells, and the intensity was significantly reduced with peptide competition in all cell lines. Only the immunoreactive band at 170 kDa was constant across the different cell lines, whereas, nonspecific bands were found to vary between the different cell lines (figure 6.11). Five potential NHS isoforms can be identified on the database with predicted molecular weights ranging from 160-196 kDa. It is not known which isoform is represented by the immunoreactive band resolving at approximately 170 kDa in these cell lines. Although these results imply that each of these five cell lines contain a predominant NHS isoform, it cannot be ruled out that other NHS isoforms are present.

MTLn3, Caco-2, and MDCK are all epithelial cell lines derived from rat, human, and dog, respectively (table 2.5). COS-7 cells are a fibroblast cell line derived from monkey (table 2.5), whilst CHO cells are derived from Chinese Hamster ovaries. Since the anti-NHS anti-peptide antibody was raised to a highly conserved peptide at the C-terminus of NHS, detection of NHS in other species was expected. COS-7 and MDCK cells are derived from kidney, a tissue in which expression of the *NHS* transcript was demonstrated (section 4.2.2.1, figure 4.16). EST sequences on the database, aligning with the *NHS* gene include transcripts derived from colon, the tissue from which Caco-2 cells are derived. Although expression of *NHS* in the mammary gland and ovaries have not yet been confirmed, endogenous NHS was detected in CHO and MTLn3 cells in this thesis.

As described previously for MTLn3 cells (Bailly *et al.*, 1999), the Arp2/3 complex was found to localise to the leading edge of the lamellipods in stimulated MTLn3 cells (figure 6.12, A and B). Localised polymerisation of actin at the plasma membrane results in protrusions such as lamellipodia. Staining of the actin cytoskeleton with Alexa Fluor 488 phalloidin revealed actin enriched at the leading edge of the lamellipodia in stimulated MTLn3 cells, colocalising with the Arp2/3 complex (figure 6.12, A and B).

Localisation of the Arp2/3 complex in unstimulated MTLn3 cells revealed more cytoplasmic staining, with some stronger staining observed at membrane ruffles (figure 6.12, C). This was consistent with previous reports of the localisation of Arp2/3 in unstimulated MTLn3 cells (Bailly *et al.*, 1999).

The cellular localisation of endogenous NHS in stimulated MTLn3 cells was reminiscent of that seen for the Arp2/3 complex. NHS was found to localise to the leading edge of lamellipods in stimulated MTLn3 cells (figure 6.13, A and B). Interestingly, the localisation of endogenous NHS in unstilmulated MTLn3 cells also revealed some membrane ruffle staining (figure 6.14), as seen in unstimulated MTLn3 cells stained for Arp2/3 (figure 6.4, Bailley *et al.*, 1999). Ruffles are formed from upward folding of lamellipodia. It is interesting that endogenous NHS should localise to the ruffles in unstimulated MTLn3 cells as this localisation pattern was observed for

the Arp2/3 complex. The staining pattern of endogenous NHS appears to closely reflect that of endogenous Arp2/3 in stimulated and unstimulated MTLn3 cells. These results are in agreement with the identified homology between the WASP protein family and the NHS protein family (section 6.2.1), which implies a role for NHS in actin cytoskeletal dynamics. To further confirm the localisation pattern at the leading edge for endogenous NHS in MTLn3 cells, the peptide synthesised to generate the anti-NHS antibody SG1703 (section 6.2.2.1) could be used to block the endogenous staining observed.

A notable difference in the localisation of endogenous NHS compared to that observed for the Arp2/3 complex in this study was the occasional decoration of some stress fibres with the anti-NHS antibody (figures 6.13, B and 6.14, B). This was not observed with the pre-immune serum (figure 6.13, C), but did occur in both stimulated and unstimulated MTLn3 cells (figures 6.13, B, and 6.14, B). This suggests that NHS may bind actin directly or through an interaction with another protein.

Taken together, the data imply a function for the NHS protein in actin cytoskeletal dynamics. These results also suggest a function for the NHS protein in activating and stimulating the Arp2/3 complex. In the future it will be worth directly testing whether the NHS protein binds the Arp2/3 complex and actin.

- Chapter Seven -

Overall discussion and future work

Cataract is a classic example of a genetically heterogenous disease. It is defined by an opacity of the lens, which can dramatically reduce visual acuity. Cataract can vary considerably in its morphology and can either be static or progressive in nature. Inherited cataract can be non-syndromic, in which case it frequently exhibits an autosomal mode of inheritance, or it can form part of a syndrome. Syndromic forms of cataract include the Lowe and Nance-Horan syndromes. The mapping of a five-generation family with non-syndromic X-linked congenital cataract (CXN, Francis *et al.*, 2002) is one of two reported cases of X-linked inherited cataract. Autosomal recessive modes of inheritance have been reported, but are more frequently seen in consanguineous families.

Genetic studies have contributed to the idea that genes involved in early onset cataract are also implicated in age-related cataract. In particular, mutations in some genes (*MIP* and *γC-crystallin*) result in progressive cataracts (Francis *et al.*, 2000; Ren *et al.*, 2000), whilst familial adult onset pulverulent cataracts has been linked to the CAAR locus (Heon *et al.*, 2001). It is becoming increasing clear therefore, that mutations with detrimental effect in certain genes results in congenital cataract, whilst mutations with less severe functional consequences in the same genes may contribute to age-related cataract in a more complex multifactorial fashion. In the era of the HapMap (section 1.7.2) it will be interesting to begin to determine which genes are implicated in susceptibility to age-related cataract.

The cataract phenotype observed in affected males of the X-linked congenital cataract family (CXN), was of a total opacity of the lens. Total lens opacities usually require surgery from an early age. In addition to the cataract phenotype, a few affected males of the CXN family were found to suffer from cardiac abnormalities, such as ventriculoseptal defect (VSD). Whether or not the cardiac anomalies completely segregated with the cataract phenotype in this family is unknown. The CXN disease locus was originally mapped to a 3.5 Mb interval on chromosome Xp22.13 – 22.2 flanked by markers DXS999 and DXS9902. Interestingly, the CXN disease locus was found to reside within the disease locus for Nance-Horan syndrome (NHS), which was mapped to a 30 cM interval on Xp22.13-22.31 flanked by markers DXS451 and DXS143 (Stambolian *et al.*, 1990; Zhu *et al.*, 1990; Bergen *et al.*, 1994).

Nance-Horan Syndrome (MIM 302350) is a rare X-linked disease with affected males exhibiting a total opacity of the lens at birth. The total lens opacity phenotype observed in affected males with NHS has also been associated with microcornea and/or microphthalmia, distinctive dental anomalies including supernumerary incisors and crown shaped permanent teeth, characteristic facial features (anteverted pinnae, long face, prominent nasal bridge and nose) and mental retardation in approximately a third of cases. Carrier females usually have cataracts confined to the posterior Y sutures and can exhibit widely spaced cone or screwdriver shaped teeth (Bixler *et al.*, 1984; Seow *et al.*, 1985).

The mapping of CXN and NHS to the short arm of the human X chromosome together with the total lens opacities observed in both, implied that the two diseases may have been allelic. Strengthening this argument was the mapping of X-linked congenital cataract in the mouse (Xcat) to the distal portion of the mouse X chromosome. The Xcat mouse exhibited a total opacity of the lens at birth in affected males and homozygous females, and a variable cataract phenotype in heterozygous females (Stambolian *et al.* 1994). The distal portion of the mouse X chromosome, to which the Xcat locus was mapped, is syntenic with the portion of the human X chromosome containing the CXN and NHS disease loci. Thus, the Xcat mouse is a potential model for CXN and/or NHS.

In this study, the causative gene for congenital cataract in the CXN family was investigated. Segregation of disease in the CXN family with markers on Xp22.13-22.2 was confirmed as part of this thesis through genotyping 46 known microsatellite markers spanning the human X chromosome. An attempt was then made to refine the CXN disease interval through the identification of novel microsatellite markers within the CXN disease locus. Novel microsatellite markers S4 and S5 were found to be recombinant in individual III.3, and in subsequent generations of the CXN family. As a result, the CXN disease interval was successfully refined to approximately 3.2 Mb leading to the exclusion of three genes (*CDKL5*, *RS1*, and *PPEF1*).

Due to the possible allelism between CXN and NHS, four British families with NHS were recruited to the study. At the time, the interval for NHS had also been refined to approximately 1.3 Mb on Xp22.13, flanked by markers DXS1995 and DXS999 (Toutain *et al.*, 2002). In addition to the refinement, five known candidate genes (*SCML1*, *SCML2*, *CDKL5*, *RS1*, and *PPEF1*) within the refined NHS disease locus had been screened and excluded as disease causative for NHS (Toutain *et al.*, 2002). The data described in this thesis and that presented by Toutain and colleagues (2002) still pointed towards possible allelism between CXN and NHS. Firstly, the refinement of the CXN locus occurred within the newly refined disease interval for NHS leaving an overlap of approximately 900 Kb of genomic sequence between the CXN and NHS disease loci. Secondly, refinement of the CXN disease locus resulted in the exclusion of *CDKL5*, *RS1*, and *PPEF1* as candidates for disease in this family. This closely reflected the report by Toutain and colleagues (2002) who screened and excluded these three genes as candidates for NHS.

Three candidate genes (*TL1*, *RBBP7*, and *CX43-L*) were selected for screening in the CXN family and NHS families based on expression in relevant tissues, homologies to known proteins or classes of proteins, and their positions within the disease intervals for CXN and NHS. No sequence changes were identified in any of the three genes resulting in their exclusion as candidates for disease in the CXN and NHS families. At this point, all known genes residing in the region of overlap between the recently refined CXN and NHS disease loci had been screened and excluded for either CXN or NHS (this thesis; Toutain *et al.*, 2002).

A novel gene prediction supported by mRNAs and EST sequences was then identified through bioinformatic analysis of the region of overlap between the NHS and CXN loci. Upon screening this novel gene truncating mutations were identified in the large exon of two of the families with NHS. However, no mutation was identified in the CXN family or the remaining two families with NHS, implying that the full coding region of the novel gene may not have been fully identified. An article was then published detailing findings of a novel gene (*NHS*) harbouring truncating mutations in 5 families with NHS (Burdon *et al.*, 2003). As anticipated, the novel gene in which truncating mutations were identified in two families with NHS in this study, was indeed the *NHS* gene reported by Burdon and colleagues (2003). Furthermore, Burdon and colleagues (2003) revealed a different 5' end of the *NHS* gene when compared to the 5' end identified in this thesis. The alternative 5' exon (exon 1a) of the *NHS* gene, resided approximately 350 Kb upstream of exon 2 (Burdon *et al.*, 2003). It became apparent therefore, that bioinformatic analysis alone would have failed to detect this exon, and that RT-PCR as done by Burdon *et al.*, 2003, was necessary to detect this alternative transcript.

The *NHS* gene was reported to be transcribed as two isoforms (A and B; Burdon *et al.*, 2003), although there are now five putative *NHS* isoforms on the database (isoforms A-E). Isoform A, referred to as the major isoform of the *NHS* gene (Burdon *et al.*, 2003), contained a 5' exon approximately 350 Kb upstream of exon 2. Isoform B consisted of an alternative 5' exon (exon 1b) with predicted translation of the protein product beginning at the first methionine of exon 4 (Burdon *et al.*, 2003). Both exon 1a and exon 1b were screened in the remaining two families with NHS, and the CXN family in this study. Truncating mutations in exon 1a were identified in the remaining two NHS families, but no sequence change was identified in either exon 1a or 1b in the CXN family.

Ramprasad and colleagues (2005) have since reported the identification of a truncating mutation in exon 1a of the *NHS* gene in a family with NHS. The cumulative results of the mutations identified in the *NHS* gene to date (this thesis, Burdon *et al.*, 2003, and Ramprasad *et al.*, 2005) have highlighted the loss of isoform A is sufficient to cause disease in families with NHS. There does not, however, appear to be any genotype/phenotype correlation as yet.

Truncating mutations within exon 1a effect isoform A only, whereas truncating mutations identified in exons 2, 3, and 4-8 are predicted to affect all *NHS* isoforms except isoform B where translation begins in exon 4. One might predict therefore, that a truncating mutation in an exon predicted to affect all isoforms of the *NHS* gene (exons 4-8), would lead to a more severe phenotype. One family reported by Burdon and colleagues (2003) opposes this idea. Burdon *et al.*, (2003) identified a truncating mutation in exon 6 of a family with NHS. Only two affected males within this family exhibited developmental delay, suggesting genetic modifiers may affect phenotype.

Orthologs of the NHS protein were sought and identified in mouse (Mus *musculus*), rat (Rattus *norvegicus*), chimpanzee (Pan *Troglodytes*), dog (Canis *familaris*), orangutan (Pongo *pygmaeus*, partial), chicken (Gallus *gallus*, partial), cow (Bos *Taurus*, partial), frog (Xenopus *laevis*, partial), zebrafish (Danio *rerio*, partial), and pufferfish (Tetradon *nigroviridis*, partial). Alignment of the orthologous protein sequences revealed a high degree of conservation. Furthermore, deduced genomic structures of full length NHS orthologues revealed a high degree of conservation in genomic organisation. The presence of an NHS ortholog in the mouse provides an excellent candidate for the Xcat mouse model.

Of the NHS families reported by Burdon and colleagues (2003), one family did not harbour a mutation in the coding exons of the *NHS* gene. The possibility that NHS may be heterogenous should not be ruled out. The closely related X-linked developmental syndrome Oculofaciocardiodental syndrome (MIM 300166) results from mutations in the *BCOR* gene (Ng *et al.*, 2004). The *BCOR* gene would make an ideal candidate gene for families with apparent NHS for which no mutation in the NHS gene is identified and for which linkage to Xp22.13 has not been confirmed. Recently, a family with NHS was mapped to the long arm of chromosome X (Xq) by Burdon and colleagues (personal communication). It is not known whether the family mapping to Xq is the same family for which a *NHS* gene mutation was not identified (Burdon *et al.*, 2003). The identification of *NHS* paralogs described in this thesis makes *NHSL2* (mapping to Xq13.1) an ideal candidate for the NHS family linked to Xq.

Expression analysis of the *NHS* gene described in this thesis revealed the presence of transcripts in fetal brain, kidney, lung, and thymus. A detailed investigation into the

expression of the *NHS* gene was also carried out by Burdon and colleagues (2003). *NHS* transcripts were detected in human adult and fetal brain, lens, retina, retinal pigment epithelium, placenta, lymphocytes and fibroblasts. Developmental expression of *Nhs* in the mouse brain was also investigated revealing varying levels of expression of *Nhs* from E17.5 through to 12 months. *In situ* hybridisation detected expression of mouse *Nhs* in midbrain, lens, retina, tooth primordia, the olfactory epithelium, the whisker follicles, the choroid plexus, and heart. To complement these data, a *lacZ* reporter-gene cassette was inserted into intron 1a of isoform A of the *Nhs* gene. This allowed Burdon and colleagues to examine the expression of the *Nhs* gene driven by the *Nhs* promoter. Strong expression was detected in the developing lens and relatively high expression was also detected in some tissues of the developing and early postnatal brain. Expression in the olfactory bulbs, olfactory epithelium and the limbic system were most prominent and expression in the heart, apical ectoderm ridge and the olfactory tubercle was observed (Burdon *et al.*, 2003). No phenotype in the lacZ mouse (i.e. lens opacities) was reported by Burdon and colleagues (2003).

To date, no mutation within the *NHS* gene has been identified in the CXN family. It is still possible that CXN and NHS are allelic and an attempt to identify alternate 5' ends or cryptic exons of the *NHS* gene may facilitate the identification of a mutation in the CXN family. As with the alternative 5' exon approximately 350 Kb upstream of exon 2 of the *NHS* gene, bioinformatics alone will not be sufficient to identify alternative 5' and cryptic exons. RT-PCR analysis and 5' RACE of lens and heart mRNA should be carried out. Upstream and other regulatory regions of the gene should also be considered as potential mutation sites leading to disease. Identification of *NHS* gene expression in the heart (Burdon *et al.*, 2003) provides further support for the *NHS* gene to be disease causative in the CXN family. In addition the Xcat mouse maps to a region of the mouse X chromosome syntenic with the interval to which the CXN and NHS loci were mapped. Thus, the Xcat mouse may prove to be a model for disease in the CXN family and/or NHS families. In the advent that CXN and NHS prove not to be allelic, new candidate genes should be sought and screened in the CXN family. It should be noted that affected individuals of the CXN family do not exhibit typical NHS features (i.e. dental anomalies and facial dysmorphism). If the *NHS* gene is disease causative for CXN, then the variation in phenotype may be a result of genetic modifiers.

Through this study, the *NHS* gene has been found to form part of a new gene family. The *NHS* paralogs: *NHSL1*, *NHSL2*, and *NHSL3*, reside on chromosomes 6q24.1, Xq13.1, and 1p35.1, respectively. Each of the *NHS* paralogs has a strikingly similar genomic structure and their protein products share between 22-28 % sequence identity. Initial expression analysis of *NHSL1* and *NHSL2* in a fetal cDNA panel revealed both to be more widely expressed than the *NHS* gene. NHSL3 appears to be the most distantly related of the family. Alignment of the protein sequences for each member of the *NHS* gene family has revealed four conserved homology domains of unknown function. An additional five homology domains have been identified between the more closely related NHS, NHSL1, and NHSL2 protein sequences. Currently, there are no clues as to the function of the various conserved homology domains in the NHS protein family apart from homology domains 1 and 8 (both absent from NHSL3). Homology domain 1 present in NHS, NHSL1, and NHSL2, was found to share sequence identity with the WAVE homology domain (WHD), whilst homology domain 8 is predicted to contain a putative actin-binding (WH2) domain.

The WHD is located at the N-terminus of the WAVE proteins 1-3. WAVE proteins form part of a larger family known as the WASP protein family, which includes WAVE1-3, WASP, and N-WASP. Each member of the WASP protein family is a known actin-binding protein and activator of the Arp2/3 complex, a seven-subunit complex that initiates actin polymerisation at the leading edge of motile cells resulting in the formation of protrusive structures such as lamellipodia and filopodia. Identification of a putative WHD at the N-terminus of members of the NHS protein family (NHS, NHSL1, and NHSL2), prompted a search for additional regions of homology shared between the NHS and WASP protein families. This led to the identification of at least one putative CA (central/acidic) domain, which is known from other proteins to bind the Arp2/3 complex resulting in its activation, in each member of the NHS protein family. Two putative CA domains (CA1 and CA2) were identified in the NHS protein. A putative actin-binding domain (WH2), present in the WASP protein family, was also identified in the protein sequences of NHS, NHSL1, and NHSL2 within homology domain 8. Together, the WH2 and CA domains of the WASP protein family facilitate binding to an actin monomer and the Arp2/3 complex (resulting in its activation), bringing the two into close proximity.

Each member of the WASP protein family also contains a proline-rich region, reported to bind SH3 domain containing proteins such as profilin. With the exception of NHSL2, each member of the NHS protein family also contains a proline-rich region. An additional domain common to all members of the WASP protein family and found in all members of the NHS protein family was a basic region. The basic region enhances the ability of WASP and N-WASP to activate the Arp2/3 complex through binding the phosphoinositide, PIP2 (Higgs and Pollard, 2000; Rohatgi *et al.*, 2000). For WAVE2, the basic region was recently reported to bind PtdIns(3,4,5)P_3, and is essential for the formation of lamellipodia (Oikawa *et al.*, 2004).

Therefore, the NHS protein family appear to share homology with the WASP protein family and are more closely related to the WAVE proteins. The presence of putative WHD, CA, and WH2 domains implies that the NHS protein family may be able to bind actin and activate the Arp2/3 complex.

The function of the WHD in the WAVE proteins was unknown until recently. A sequence motif within the WHD has now been identified as an Abelson interactor (Abi)-binding domain (Echarri *et al.*, 2004; Leng *et al.*, 2005). Binding of Abi1 to the WHD of WAVE 1 and WAVE 2, results in the translocation of the WAVE proteins to the leading edge of motile cells (Echarri *et al.*, 2004; Leng *et al.*, 2005).

Unlike Abi2, which is highly expressed in the lens, Abi1 is not expressed in the lens (Grove *et al.*, 2004). This is of particular interest with regards to the putative Abi-binding domain in the NHS protein family. Should the putative Abi-binding domain in NHS prove to be functional, its binding partner in the lens is likely to be Abi2. The level of homology between the WHD of the WAVE proteins 1-3 and the N-terminus of the NHS, NHSL1, and NHSL2 implies that the NHS protein family may contain an Abi-binding domain. Since Abi1 has been reported to bind and translocate WAVE1 and WAVE2 to the leading edge of lamellipodia (Echarri *et al.*, 2004; Leng *et al.*, 2005), the NHS proteins may also bind an Abi protein and be translocated to lamellipodia. The variation in amino acid sequence over the putative Abi-binding domains in NHS and NHSL2 when compared to the WAVE proteins, may result in their specificity for an alternative Abi family member.

The sub-cellular localisation of NHS in this study supports the notion that NHS may bind an Abi protein through its putative WHD and/or actin. NHS was found to localise to the leading edge of lamellipods in stimulated MTLn3 cells. The binding of NHS to an Abi family member may result in its translocation to the leading edge of motile cells.

To determine whether the putative WHD in the NHS protein family is functional and binds one of the Abi protein family members, co-immunoprecipitation and GST-pull down assays should be performed.

Interestingly, a recent paper reported the generation of an Abi2 null mutant mouse, which developed abnormal phenotypes in the eye and brain (Grove *et al.*, 2004). The eyes of adult $Abi2^{-/-}$ mice were noticeably smaller than those of wild-type mice, and appeared to be a result of defective lens development. Primary lens fibres at E14.4 were found to have completely filled the lens vesicle in both wild-type and Abi2 null mice. However, the orientation and migration of secondary lens fibres in Abi2 null mice was defective at E16.5 and P1. In addition, the anterior and posterior sutures failed to develop in Abi2 null mice. At P1 the lens was found to be very deformed and the lens capsule to be substantially thicker. Failure for the sutures to form resulted in rupture of the posterior region of the lens and subsequent release of primary lens fibres (Grove *et al.*, 2004). Grove and colleagues further investigated the defects observed in secondary lens fibre migration. They localised the Abi2 protein in wild-type lenses to the tips of migrating secondary fibres. In particular they found Abi2 to localise at the transition zone where newly differentiated fibres begin migration along the epithelial cell-fibre interface (EFI). It is at the EFI that secondary lens fibres form adherens junctions with anterior epithelial cells. They also found Abi2 to localise with β catenin along secondary lens fibre cell-cell borders. The catenins, α and β, provide a link between cadherins and the actin cytoskeleton. Furthermore, down regulation of Abi1 and Abi2 in HeLa epithelial cells, prevented formation of adherens junctions (Grove *et al.*, 2004). Thus, Abi2 appears to play a role in adherens junction formation.

Neuronal abnormalities were also observed in Abi2 null mice. Abnormalities in the organisation of the neocortex, and mild distortion in the organisation of the hippocampus were observed in adult Abi2 null mice. Dendritic spine morphology and density were also abnormal in Abi2 null and heterozygous mice, and behavioural tests

conducted demonstrated deficits in the short and long term memory of Abi2 null mice (Grove *et al.*, 2004).

The overlapping phenotype between NHS and Abi2 null mice, and the potential Abi-binding domain in the NHS protein along with observed expression of both NHS and Abi2 in the lens, suggests a possible role for NHS in adherens junction formation. It has already been reported that polymerisation of actin is the driving force behind the formation of adherens junctions (Vasioukhin *et al.*, 2000), and therefore, Abi2 may bring NHS into the vicinity of the Arp2/3 complex for localised actin polymerisation and subsequent generation of adherens junctions.

Investigation of the potential co-localisation of endogenous NHS and junctional proteins such as cadherins and α- and β-catenin in MTLn3 cells or another epithelial cell line expressing NHS (i.e. MDCK), should be carried out in the future using marker antibodies and the anti-NHS antibody described here. This will determine whether or not NHS localises to adherens junctions. The junctional proteins cadherins and the α- and β-catenin proteins are also expressed in lens fibre cells, and determining whether NHS localises to adherens junctions in these lens specific cells will be important, as cataract is a hallmark of Nance-Horan syndrome.

Should NHS be found to localise to adherens junctions, its role in adherens junction formation should be investigated through RNAi. By transfecting cells with NHS-specific siRNAs the affect on adherens junction formation could be determined through immunostaining of an junctional protein, such as β-catenin.

Recent reports further support the idea that the NHS protein is related to the WAVE proteins and has a function in actin polymerisation. Firstly, the WAVE1-associated GTPase-activating/Mental disorder-associated GAP protein (WRP/MEGAP), which terminates Rac signalling in neurons (Soderling *et al.*, 2002; Endris *et al.*, 2002), has been identified as the underlying cause in 3p (-) syndrome (Endris *et al.*, 2002). Features of the 3p (-) syndrome include hypotonia and severe mental retardation. These features are analogous with the symptoms observed in the affected male from NHS family 3 in this study (severe developmental delay and hypotonia). Secondly, targeted disruption of the *Wave1* gene in mouse resulted in reduced anxiety, sensorimotor

retardation, and deficits in hippocampal-dependent learning and memory. These symptoms reflect those observed in patients with 3p (-) syndrome (Soderling *et al.*, 2003), and approximately a third of the cases of NHS suffer mental retardation.

The possible link between the Arp2/3 complex and actin and the NHS protein was investigated. MTLn3 cells were confirmed to express endogenous NHS using an anti-NHS antibody described in this thesis. These cells have been used extensively by my collaborator Dr M. Bailly, to study the actin cytoskeleton and localisation of the Arp2/3 complex upon stimulation to induce lamellipod formation. To confirm the localisation pattern previously observed for Arp2/3 in MTLn3 cells, an anti-p34 subunit antibody was used to immunostain stimulated and unstimulated MTLn3 cells.

The Arp2/3 complex localised to the leading edge of lamellipodia in stimulated MTLn3 cells, as previously reported (Bailly *et al.*, 1999). Interestingly, the subcellular localisation of endogenous NHS in stimulated MTLn3 cells closely resembled that of the Arp2/3 complex. Endogenous NHS was found to localise at the leading edge of the lamellipods. Both the Arp2/3 complex and the NHS protein were found to localise to membrane ruffles in unstimulated MTLn3 cells.

Endogenous NHS was also localised to actin stress fibres in stimulated and unstimulated MTLn3 cells. This sub-cellular localisation pattern was not observed for the Arp2/3 complex.

The localisation of endogenous NHS to the leading edge of lamellipods and the identification of putative WH2 and CA domains, suggest that NHS may be a nucleation promoting factor for the Arp2/3 complex.

Other putative domains identified within the NHS protein which are important to the function of the WAVE proteins were the proline-rich and basic regions. Proline-rich regions of the WAVE proteins are binding sites for SH3 domain containing proteins such as profilin. Potential binding of profilin could be addressed for the NHS protein through co-immunoprecipitation and GST pull down assays.

PtdIns(3,4,5)P_3 was recently shown to bind the WAVE2 protein and is important for lamellipod formation (Oikawa *et al.*, 2004), whilst the basic region in WASP and N-

WASP binds PIP2 enhancing their ability to activate the Arp2/3 complex (Higgs and Pollard, 2000; Rohatgi *et al.*, 2000). Sub cellular localisation of NHS to the leading edge of MTLn3 cells in this study supports the idea that the basic region of NHS may bind either PtdIns(3,4,5)P_3, PIP2 or another phosphoinositide. Binding of WAVE2 to PtdIns(3,4,5)P_3 was determined through enzyme-linked immunosorbent assays (ELISA, Oikawa *et al.*, 2004), which could be applied to the NHS protein.

To determine whether the putative CA1 domain in NHS was functional a competitive actin polymerisation assay was performed. Unfortunately, the results obtained from the competitive actin polymerisation assays were inconclusive. Cloning a larger fragment of the NHS protein or the entire protein containing the putative WH2 and CA domains for a non-competitive actin polymerisation assay, would be a good approach to determine whether the NHS protein is a nucleation promoting factor of the Arp2/3 complex.

The Arp2/3 complex polymerises actin at the side of pre-existing actin filaments resulting in formation of Y-shaped branches (Pollard *et al.*, 2000). Such branches are commonly seen in actin meshworks of leading edge lamellipodia in motile cells (Svitkina and Borisy, 1999; Bailly *et al.*, 1999). Should the NHS protein prove to activate the Arp2/3 complex, it will be important to determine whether its activation results in the characteristic branched filaments. This can be achieved through *in vitro* polymerisation of actin in the presence of an NHS construct and Arp2/3 complex and labelled phalloidin, which can be fluorescently visualised. Control experiments can also be run in parallel consisting of actin monomers on their own (which leads to long unbranched filaments) and actin in the presence of the Arp2/3 complex and a known nucleation promoting factor construct such as WASP. Disruption of branched actin networks could also be determined through use of mutant NHS constructs. If NHS does stimulate the Arp2/3 complex thus resulting in branched actin filaments, then the NHS protein can be localised in the branched actin networks through high resolution replica electron microscopy (Bailly *et al.*, 1999).

The ability of the NHS protein to bind G-actin and the Arp2/3 complex could be determined through GST-pull down assays. NHS constructs immobilised on glutathione-sepharose beads are incubated with monomeric actin or the Arp2/3

complex, washed, and the bound protein eluted for Western blot analysis using commercially available anti-actin antibodies and the anti-p34 antibody used in this thesis. Use of mutant NHS constructs will enable key residues involved in binding monomeric actin and the Arp2/3 complex to be identified.

The affect of knocking down NHS expression in cell lines expressing endogenous NHS (for example, MTLn3, MDCK, Caco-2) through RNAi should be investigated to see what effect reduced or absent NHS expression has on the actin cytoskeleton, and as previously mentioned, on adherens junction formation.

It will be important to generate additional antibiodies against the NHS protein, in particular against specific isoforms of NHS such as isoform A, which appears to be sufficient to cause disease in families with Nance-Horan syndrome and is predicted to contain the WHD. Since Nance-Horan Syndrome is a developmental disorder, antibodies raised against the NHS protein will provide a valuable resource for examining the spatiotemporal expression pattern during development, and subcellular localisation of the various NHS isoforms.

Affected males with Nance-Horan syndrome exhibit total cataract whilst heterozygous females generally have sutural cataracts. These cataract phenotypes described in patients with Nance-Horan syndrome may result from aberrant migration of the lens fibre cells and failure to fully form adherens junctions. During lens development, secondary lens fibres elongate and differentiate forming sutures at points where they come into apposition (section 1.2.1). The failure of secondary lens fibres to form adherens junctions with neighbouring secondary lens fibres may result in cataract formation at the points of apposition leading to a characteristic Y-sutural cataract phenotype observed in heterozygous females. In total cataract, all lens fibre cells are affected. Since affected males only have one copy of the *NHS* gene andno functional NHS, adherens junctions may not form in both primary and secondary lens fibres providing an explanation for the more severe cataract phenotype observed. Defective adherens junction formation is one proposed disease mechanism for cataract in patients with Lowe syndrome (Faucherre *et al.*, 2005).

How mutations in the *NHS* gene result in the dental anomalies common to Nance-Horan syndrome is unclear. Currently there does not appear to be any genotype/phenotype correlation between potential loss of certain or all NHS isoforms and the dental anomalies observed in NHS families. Genes mutated in human dental disorders include transcription factors such as *BCOR* (Ng *et al.*, 2004), and the signalling molecule ectodysplasin (EDA). Disruption of EDA, its receptor EDAR, and the intracellular adaptor protein EDARADD in mice results in deviances from the normal number of teeth formed (Sofaer, 1969; Sofaer, 1977; Headon *et al.*, 2001), highlighting that defects in signalling pathways result in dental anomalies. NHS may act as a scaffolding protein downstream of signalling cascades involved in tooth development, which is supported by the identification of *NHS* expression in tooth primordial (Burdon *et al.*, 2003) and the dental anomalies that form part of the Nance-Horan syndrome phenotype. In rare cases, supernumerary teeth have been observed following loss of Eda or Edar and interestingly, supernumerary incisors form part of the dental anomalies seen in Nance-Horan Syndrome patients.

If the *Nhs* gene is disease causative in the Xcat mouse, the absence of dental anomalies in Xcat mice is not entirely surprising. Compared to humans who have more tooth types (canines and premolars), mice only have one set of teeth, which develop in a manner similar to that of human deciduous teeth (milk teeth; Tucker and Sharpe, 2004). Thus, the mouse is not a particularly good model for human dental disorders.

The NHS protein may also have a role in modulating gene expression (figure 7.1). Four monopartite nuclear localisation signals have been identified in NHS, implying that NHS could be imported into the nucleus. Disruption of NHS therefore, may alter the regulation of gene expression resulting in the tooth abnormalities observed in patients with NHS. Furthermore, craniofacial development is dependent on an intricate pattern of gene expression and mutations affecting proteins that modulate gene expression can lead to abnormal craniofacial development. If NHS is imported into the nucleus and modulates gene expression, then the facial dysmorphism observed in patients with NHS may arise from loss of gene modulation in development by the NHS protein.

To summarise, the NHS protein may act as a scaffolding protein bringing together various binding partners into one complex (as described for the WAVE proteins; Higgs

and Pollard, 1999) and activating the Arp2/3 complex for localised actin polymerisation (figure 7.1). The Drosophila GUKH protein is also a scaffolding protein and the homology identified between NHS and the Drosophila GUKH protein (Katoh and Katoh, 2004) may be due to domains that are common to scaffolding proteins. In addition to the sequence similarity between the WASP protein family and the NHS protein family, homology domains have also been identified in the NHS protein family for which a potential function has not been attributed. Further, Burdon and colleagues (2003) reported four monopartite nuclear localisation signals (NLS) in the NHS protein, although two do not appear to be functional (this study, section 6.2.2.3.2). Bipartite nuclear localisation signals have been identified in other members of the NHS protein family (except NHSL3), which may localise these proteins to the nucleus under certain physiological conditions.

Clearly the research presented in this thesis presents exciting avenues to follow in order to elucidate the potential functions of the WASP protein family related domains, and the additional homology domains within the NHS protein family. This will aid in gaining a greater insight into the disease mechanism underlying Nance-Horan syndrome.

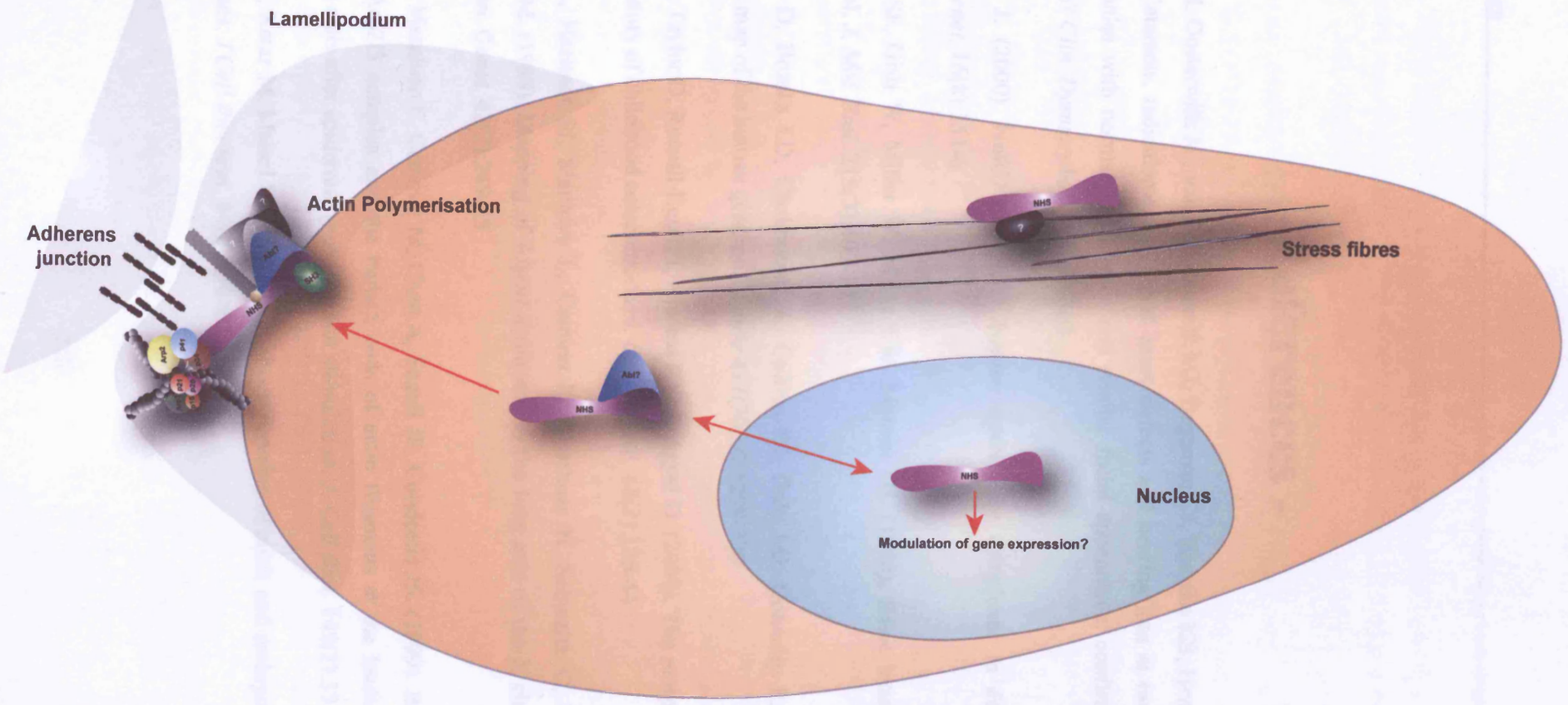


Figure 7.1 Proposed model for the role of NHS

In response to extracellular stimuli, NHS could bind a member of the Abi protein family and an SH3 domain containing protein, resulting in NHS being translocated to the plasma membrane (leading edge) where it binds actin and activates the Arp2/3 complex leading to the formation of a lamellipod. Upon certain stimuli, NHS may be imported into the nucleus where it modulates gene expression. The NHS protein could localise to actin stress fibres where it acts as a scaffold, bringing together signalling molecules and the cytoskeleton.

-References -

Aalfs CM, Oosterwijk JC, van Schooneveld MJ, Begeman CJ, Wabeke KB, Hennekam RC. (1996). Cataracts, radiculomegaly, septal heart defects and hearing loss in two unrelated adult females with normal intelligence and similar facial appearance: confirmation of a syndrome? *Clin. Dysmorphol.* 5(2):93-103

Ahringer J. (2000). NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet.* 16(8):351-6

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215(3):403-10

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. (2005). A haplotype map of the human genome. *Nature* 437(7063):1299-320

Amaya L, Taylor D, Russell-Eggitt I, Nischal KK, Lengyel D. (2003). The morphology and natural history of childhood cataracts. *Surv. Ophthalmol.* 48(2):125-44

Atkin CL, Hasstedt SJ, Menlove L, Cannon L, Kirschner N, Schwartz C, Nguyen K, Skolnick M. (1988). Mapping of Alport syndrome to the long arm of the X chromosome. *Am. J. Hum. Genet.* 42(2):249-55

Bailly M, Macaluso F, Cammer M, Chan A, Segall JE, Condeelis JS. (1999). Relationship between Arp2/3 complex and the barbed ends of actin filaments at the leading edge of carcinoma cells after epidermal growth factor stimulation. *J. Cell Biol.* 145(2):331-45

Bagchi M, Katar M, Maisel H. (2002) Heat shock proteins of adult and embryonic human ocular lenses. *J Cell Biochem.* 84(2):278-84.

Bear JE, Rawls JF, Saxe CL, III. (1998). SCAR, a WASP-related protein, isolated as a suppressor of receptor defects in late Dictyostelium development. *J. Cell Biol.* 142(5):1325-35

Beaumont C, Leneuve P, Devaux I, Scoazec JY, Berthier M, Loiseau MN, Grandchamp B, Bonneau D. (1995). Mutation in the iron responsive element of the L ferritin mRNA in a family with dominant hyperferritinaemia and cataract. *Nat. Genet.* 11(4):444-6

Benachenhou N, Massy I, Vacher J. (2002). Characterization and expression analyses of the mouse Wiskott-Aldrich syndrome protein (WASP) family member Wave1/Scar. *Gene* 290(1-2):131-40

Bergen AA, ten Brink J, Schuurman EJ, Bleeker-Wagemakers EM. (1994). Nance-Horan syndrome: linkage analysis in a family from The Netherlands. *Genomics* 21(1):238-40

Berridge MJ, Irvine RF. (1989). Inositol phosphates and cell signalling. *Nature* 341(6239):197-205

Berry V, Francis P, Kaushal S, Moore A, Bhattacharya S. (2000). Missense mutations in MIP underlie autosomal dominant 'polymorphic' and lamellar cataracts linked to 12q. *Nat. Genet.* 25(1):15-7

Berry V, Francis P, Reddy MA, Collyer D, Vithana E, MacKay I, Dawson G, Carey AH, Moore A, Bhattacharya SS, Quinlan RA. (2001). Alpha-B crystallin gene (CRYAB) mutation causes dominant congenital posterior polar cataract in humans. *Am. J. Hum. Genet.* 69(5):1141-5

Biesova Z, Piccoli C, Wong WT. (1997). Isolation and characterization of e3B1, an eps8 binding protein that regulates cell growth. *Oncogene* 14(2):233-41

Birney E, Bateman A, Clamp ME, Hubbard TJ. (2001). Mining the draft human genome. *Nature* 409(6822):827-8

Bixler D, Higgins M, Hartsfield J, Jr. (1984). The Nance-Horan syndrome: a rare X-linked ocular-dental trait with expression in heterozygous females. *Clin. Genet.* 26(1):30-5

Blanchoin L, Amann KJ, Higgs HN, Marchand JB, Kaiser DA, Pollard TD. (2000). Direct observation of dendritic actin filament networks nucleated by Arp2/3 complex and WASP/Scar proteins. *Nature* 404(6781):1007-11

Bogdan S, Stephan R, Lobke C, Mertens A, Klambt C. (2005). Abi activates WASP to promote sensory organ development. *Nat. Cell Biol.* 7(10):977-84

Bompard G, Caron E. (2004) Regulation of WASP/WAVE proteins: making a long story short. *J Cell Biol.*166(7):957-62.

Breen M, Bullerdiek J, Langford CF. (1999). The DAPI banded karyotype of the domestic dog (Canis familiaris) generated using chromosome-specific paint probes. *Chromosome. Res.* 7(5):401-6

Britz-Cunningham SH, Shah MM, Zuppan CW, Fletcher WH. (1995). Mutations of the Connexin43 gap-junction gene in patients with heart malformations and defects of laterality. *N. Engl. J. Med.* 332(20):1323-9

Brown NP, Bron AJ. (1996). Development of the lens. In *Lens Disorders: A clinical manual of cataract diagnosis*, 2:4-16 pp. 4-16 pp.

Bryce NS, Clark ES, Leysath JL, Currie JD, Webb DJ, Weaver AM. (2005). Cortactin promotes cell motility by enhancing lamellipodial persistence. *Curr. Biol.* 15(14):1276-85

Bu L, Jin Y, Shi Y, Chu R, Ban A, Eiberg H, Andres L, Jiang H, Zheng G, Qian M, Cui B, Xia Y, Liu J, Hu L, Zhao G, Hayden MR, Kong X. (2002). Mutant DNA-binding domain of HSF4 is associated with autosomal dominant lamellar and Marner cataract. *Nat. Genet.* 31(3):276-8

Burdon KP, McKay JD, Sale MM, Russell-Eggitt IM, Mackey DA, Wirth MG, Elder JE, Nicoll A, Clarke MP, FitzGerald LM, Stankovich JM, Shaw MA, Sharma S, Gajovic S, Gruss P, Ross S, Thomas P, Voss AK, Thomas T, Gecz J, Craig JE. (2003). Mutations in a novel gene, NHS, cause the pleiotropic effects of nance-horan syndrome, including severe congenital cataract, dental anomalies, and mental retardation. *Am. J. Hum. Genet.* 73(6):1120-30

Callebaut I, Cossart P, Dehoux P. (1998). EVH1/WH1 domains of VASP and WASP proteins belong to a large family including Ran-binding domains of the RanBP1 family. *FEBS Lett.* 441(2):181-5

Cameron LA, Footer MJ, van Oudenaarden A, Theriot JA. (1999). Motility of ActA protein-coated microspheres driven by actin polymerization. *Proc. Natl. Acad. Sci. U. S. A* 96(9):4908-13

Carlier MF. (1990). Actin polymerization and ATP hydrolysis. *Adv. Biophys.* 26:51-73

Carlier MF, Nioche P, Broutin-L'Hermite I, Boujemaa R, Le Clainche C, Egile C, Garbay C, Ducruix A, Sansonetti P, Pantaloni D. (2000). GRB2 links signaling to actin assembly by enhancing interaction of neural Wiskott-Aldrich syndrome protein (N-WASp) with actin-related protein (ARP2/3) complex. *J. Biol. Chem.* 275(29):21946-52

Carter JM, Hutcheson AM, Quinlan RA. (1995). In vitro studies on the assembly properties of the lens proteins CP49, CP115: coassembly with alpha-crystallin but not with vimentin. *Exp. Eye Res.* 60(2):181-92

Choudhury R, Diao A, Zhang F, Eisenberg E, Saint-Pol A, Williams C, Konstantakopoulos A, Lucocq J, Johannes L, Rabouille C, Greene LE, Lowe M. (2005). Lowe syndrome protein OCRL1 interacts with clathrin and regulates protein trafficking between endosomes and the trans-Golgi network. *Mol. Biol. Cell* 16(8):3467-79

Conley YP, Erturk D, Keverline A, Mah TS, Keravala A, Barnes LR, Bruchis A, Hess JF, FitzGerald PG, Weeks DE, Ferrell RE, Gorin MB. (2000). A juvenile-onset, progressive cataract locus on chromosome 3q21-q22 is associated with a missense mutation in the beaded filament structural protein-2. *Am. J. Hum. Genet.* 66(4):1426-31

Cory GO, Garg R, Cramer R, Ridley AJ. (2002). Phosphorylation of tyrosine 291 enhances the ability of WASp to stimulate actin polymerization and filopodium formation. Wiskott-Aldrich Syndrome protein. *J. Biol. Chem.* 277(47):45115-21

Cowan CA, Henkemeyer M. (2001). The SH2/SH3 adaptor Grb4 transduces B-ephrin reverse signals. *Nature* 413(6852):174-9

Daga RR, Thode G, Amores A. (1996). Chromosome complement, C-banding, Ag-NOR and replication banding in the zebrafish Danio rerio. *Chromosome. Res.* 4(1):29-32

Dai Z, Pendergast AM. (1995). Abi-2, a novel SH3-containing protein interacts with the c-Abl tyrosine kinase and modulates c-Abl transforming activity. *Genes Dev.* 9(21):2569-82

Derry JM, Kerns JA, Weinberg KI, Ochs HD, Volpini V, Estivill X, Walker AP, Francke U. (1995). WASP gene mutations in Wiskott-Aldrich syndrome and X-linked thrombocytopenia. *Hum. Mol. Genet.* 4(7):1127-35

Devineni N, Minamide LS, Niu M, Safer D, Verma R, Bamburg JR, Nachmias VT. (1999). A quantitative analysis of G-actin binding proteins and the G-actin pool in developing chick brain. *Brain Res.* 823(1-2):129-40

Disanza A, Carlier MF, Stradal TE, Didry D, Frittoli E, Confalonieri S, Croce A, Wehland J, Di Fiore PP, Scita G. (2004). Eps8 controls actin-based motility by capping the barbed ends of actin filaments. *Nat. Cell Biol.* 6(12):1180-8

Donahue RP, Bias WB, Renwick JH, McKusick VA.(1968) Probable assignment of the Duffy blood group locus to chromosome 1 in man. *Proc Natl Acad Sci U S A.* 61(3):949-55.

Dramsi S, Levi S, Triller A, Cossart P. (1998). Entry of Listeria monocytogenes into neurons occurs by cell-to-cell spread: an in vitro study. *Infect. Immun.* 66(9):4461-8

Dressman MA, Olivos-Glander IM, Nussbaum RL, Suchy SF. (2000). Ocrl1, a PtdIns(4,5)P(2) 5-phosphatase, is localized to the trans-Golgi network of fibroblasts and epithelial cells. *J. Histochem. Cytochem.* 48(2):179-90

Drubin DG, Mulholland J, Zhu ZM, Botstein D. (1990). Homology of a yeast actin-binding protein to signal transduction proteins and myosin-I. *Nature* 343(6255):288-90

Duncan MK, Haynes JI, Cvekl A, Piatigorsky J. (1998). Dual roles for Pax-6: a transcriptional repressor of lens fiber cell-specific beta-crystallin genes. *Mol. Cell Biol.* 18(9):5579-86

Duncan MC, Cope MJ, Goode BL, Wendland B, Drubin DG. (2001). Yeast Eps15-like endocytic protein, Pan1p, activates the Arp2/3 complex. *Nat. Cell Biol.* 3(7):687-90

Echarri A, Lai MJ, Robinson MR, Pendergast AM. (2004). Abl interactor 1 (Abi-1) wave-binding and SNARE domains regulate its nucleocytoplasmic shuttling, lamellipodium localization, and wave-1 levels. *Mol. Cell Biol.* 24(11):4979-93

Eden S, Rohatgi R, Podtelejnikov AV, Mann M, Kirschner MW. (2002). Mechanism of regulation of WAVE1-induced actin nucleation by Rac1 and Nck. *Nature* 418(6899):790-3

Endo A, Ingalls TH. (1968). Chromosomes of the zebra fish. A model for cytogenetic, embryologic, and ecologic study. *J. Hered.* 59(6):382-4

Endris V, Wogatzky B, Leimer U, Bartsch D, Zatyka M, Latif F, Maher ER, Tariverdian G, Kirsch S, Karch D, Rappold GA. (2002). The novel Rho-GTPase activating gene MEGAP/ srGAP3 has a putative role in severe mental retardation. *Proc. Natl. Acad. Sci. U. S. A* 99(18):11754-9

Evangelista M, Klebl BM, Tong AH, Webb BA, Leeuw T, Leberer E, Whiteway M, Thomas DY, Boone C. (2000). A role for myosin-I in actin assembly through interactions with Vrp1p, Bee1p, and the Arp2/3 complex. *J. Cell Biol.* 148(2):353-62

Fan PD, Goff SP. (2000). Abl interactor 1 binds to sos and inhibits epidermal growth factor- and v-Abl-induced activation of extracellular signal-regulated kinases. *Mol. Cell Biol.* 20(20):7591-601

Faucherre A, Desbois P, Satre V, Lunardi J, Dorseuil O, Gacon G. (2003). Lowe syndrome protein OCRL1 interacts with Rac GTPase in the trans-Golgi network. *Hum. Mol. Genet.* 12(19):2449-56

Faucherre A, Desbois P, Nagano F, Satre V, Lunardi J, Gacon G, Dorseuil O. (2005). Lowe syndrome protein Ocrl1 is translocated to membrane ruffles upon Rac GTPase activation: a new perspective on Lowe syndrome pathophysiology. *Hum. Mol. Genet.* 14(11):1441-8

Favor J, Pretsch W. (1990). Genetic localization and phenotypic expression of X-linked cataract (Xcat) in Mus musculus. *Genet. Res.* 56(2-3):157-62

Fitch WM. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19(2):99-113

Forshew T, Johnson CA, Khaliq S, Pasha S, Willis C, Abbasi R, Tee L, Smith U, Trembath RC, Mehdi SQ, Moore AT, Maher ER. (2005). Locus heterogeneity in autosomal recessive congenital cataracts: linkage to 9q and germline HSF4 mutations. *Hum. Genet.* 117(5):452-9

Francis PJ, Berry V, Bhattacharya SS, Moore AT. (2000). The genetics of childhood cataract. *J. Med. Genet.* 37(7):481-8

Francis PJ, Berry V, Hardcastle AJ, Maher ER, Moore AT, Bhattacharya SS. (2002). A locus for isolated cataract on human Xp. *J. Med. Genet.* 39(2):105-9

Francois J. (1982). Genetics of cataract. *Ophthalmologica* 184(2):61-71

Fukuoka M, Suetsugu S, Miki H, Fukami K, Endo T, Takenawa T. (2001). A novel neural Wiskott-Aldrich syndrome protein (N-WASP) binding protein, WISH, induces Arp2/3 complex activation independent of Cdc42. *J. Cell Biol.* 152(3):471-82

Gao L, Qin W, Cui H, Feng G, Liu P, Gao W, Ma L, Li P, He L, Fu S. (2005). A novel locus of coralliform cataract mapped to chromosome 2p24-pter. *J. Hum. Genet.* 50(6):305-10

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31(13):3784-8

Gautreau A, Ho HY, Li J, Steen H, Gygi SP, Kirschner MW. (2004). Purification and architecture of the ubiquitous Wave complex. *Proc. Natl. Acad. Sci. U. S. A* 101(13):4379-83

Glaser T, Jepeal L, Edwards JG, Young SR, Favor J, Maas RL. (1994). PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nat. Genet.* 7(4):463-71

Goldschmidt-Clermont PJ, Machesky LM, Doberstein SK, Pollard TD. (1991). Mechanism of the interaction of human platelet profilin with actin. *J. Cell Biol.* 113(5):1081-9

Goode BL, Rodal AA, Barnes G, Drubin DG. (2001). Activation of the Arp2/3 complex by the actin filament binding protein Abp1p. *J. Cell Biol.* 153(3):627-34

Goodenough DA. (1980). Intercellular junctions. *Soc. Gen. Physiol Ser.* 34:167-78

Goodenough DA. (1992). The crystalline lens. A system networked by gap junctional intercellular communication. *Semin. Cell Biol.* 3(1):49-58

Gouin E, Egile C, Dehoux P, Villiers V, Adams J, Gertler F, Li R, Cossart P. (2004). The RickA protein of Rickettsia conorii activates the Arp2/3 complex. *Nature* 427(6973):457-61

Graw J. (1997). The crystallins: genes, proteins and diseases. *Biol. Chem.* 378(11):1331-48

Graw J, Loster J. (2003). Developmental genetics in ophthalmology. *Ophthalmic Genet.* 24(1):1-33

Grove M, Demyanenko G, Echarri A, Zipfel PA, Quiroz ME, Rodriguiz RM, Playford M, Martensen SA, Robinson MR, Wetsel WC, Maness PF, Pendergast AM. (2004). ABI2-deficient mice exhibit defective cell migration, aberrant dendritic spine morphogenesis, and deficits in learning and memory. *Mol. Cell Biol.* 24(24):10905-22

Hahne P, Sechi A, Benesch S, Small JV. (2001). Scar/WAVE is localised at the tips of protruding lamellipodia in living cells. *FEBS Lett.* 492(3):215-20

Hanson IM, Fletcher JM, Jordan T, Brown A, Taylor D, Adams RJ, Punnett HH, van H, V. (1994). Mutations at the PAX6 locus are found in heterogeneous anterior segment malformations including Peters' anomaly. *Nat. Genet.* 6(2):168-73

Hardcastle AJ, Thiselton DL, Zito I, Ebenezer N, Mah TS, Gorin MB, Bhattacharya SS. (2000). Evidence for a new locus for X-linked retinitis pigmentosa (RP23). *Invest Ophthalmol. Vis. Sci.* 41(8):2080-6

He W, Li S. (2000). Congenital cataracts: gene mapping. *Hum. Genet.* 106(1):1-13

Headon DJ, Emmal SA, Ferguson BM, Tucker AS, Justice MJ, Sharpe PT, Zonana J, Overbeek PA. (2001). Gene defect in ectodermal dysplasia implicates a death domain adapter in development. *Nature*. 27:414

Heon E, Priston M, Schorderet DF, Billingsley GD, Girard PO, Lubsen N, Munier FL. (1999) The gamma-crystallins and human cataracts: a puzzle made clearer. *Am J Hum Genet*. 65(5):1261-7.

Heon E, Paterson AD, Fraser M, Billingsley G, Priston M, Balmer A, Schorderet DF, Verner A, Hudson TJ, Munier FL (2001). A progressive autosomal recessive cataract locus maps to chromosome 9q13-q22. *Am J Hum Genet*. 68(3):772-7.

Heymann JB, Agre P, Engel A. (1998). Progress on the structure and function of aquaporin 1. *J. Struct. Biol.* 121(2):191-206

Higgs HN, Pollard TD. (1999). Influence of the C terminus of Wiskott-Aldrich syndrome protein (WASp) and the Arp2/3 complex on actin polymerization. *Biochemistry*. 38(46):15212-22

Higgs HN, Pollard TD. (2000). Activation by Cdc42 and PIP(2) of Wiskott-Aldrich syndrome protein (WASp) stimulates actin nucleation by Arp2/3 complex. *J. Cell Biol.* 150(6):1311-20

Hill RE, Favor J, Hogan BL, Ton CC, Saunders GF, Hanson IM, Prosser J, Jordan T, Hastie ND, van H, V. (1992). Mouse Small eye results from mutations in a paired-like homeobox-containing gene. *Nature* 355(6362):750

Huang KM, Geunes-Boyer S, Wu J, Moy C, Stambolian D. (2003). Molecular Genetics of X-Linked Cataracts: Quantitative Expression of Xcat Candidate Genes. ARVO conference. 1263/B159

Huang KM, Geunes-Boyer S, Wu S, Dutra A, Favor J, Stambolian D. (2004). Organization and annotation of the Xcat critical region: elimination of seven positional candidate genes. *Genomics* 83(5):893-901

Hudson AM, Cooley L. (2002). A subset of dynamic actin rearrangements in Drosophila requires the Arp2/3 complex. *J. Cell Biol.* 156(4):677-87

Hughes AE, Dash DP, Jackson AJ, Frazer DG, Silvestri G. (2003). Familial keratoconus with cataract: linkage to the long arm of chromosome 15 and exclusion of candidate genes. *Invest Ophthalmol. Vis. Sci.* 44(12):5063-6

Humbert P, Russell S, Richardson H. (2003). Dlg, Scribble and Lgl in cell polarity, cell proliferation and cancer. *Bioessays* 25(6):542-53

Innocenti M, Tenca P, Frittoli E, Faretta M, Tocchetti A, Di Fiore PP, Scita G. (2002). Mechanisms through which Sos-1 coordinates the activation of Ras and Rac. *J. Cell Biol.* 156(1):125-36

Innocenti M, Frittoli E, Ponzanelli I, Falck JR, Brachmann SM, Di Fiore PP, Scita G. (2003). Phosphoinositide 3-kinase activates Rac by entering in a complex with Eps8, Abi1, and Sos-1. *J. Cell Biol.* 160(1):17-23

Innocenti M, Zucconi A, Disanza A, Frittoli E, Areces LB, Steffen A, Stradal TE, Di Fiore PP, Carlier MF, Scita G. (2004). Abi1 is essential for the formation and activation of a WAVE2 signalling complex. *Nat. Cell Biol.* 6(4):319-27

Innocenti M, Gerboth S, Rottner K, Lai FP, Hertzog M, Stradal TE, Frittoli E, Didry D, Polo S, Disanza A, Benesch S, Di Fiore PP, Carlier MF, Scita G. (2005). Abi1 regulates the activity of N-WASP and WAVE in distinct actin-based processes. *Nat. Cell Biol.* 7(10):969-76

Jaffe NS, Horwitz J. (1992). Anatomy and Embryology. In *Lens and Cataract*, 1:1-1.9 pp. 1-1.9 pp.

Jakobs PM, Hess JF, FitzGerald PG, Kramer P, Weleber RG, Litt M. (2000). Autosomal-dominant congenital cataract associated with a deletion mutation in the human beaded filament protein gene BFSP2. *Am. J. Hum. Genet.* 66(4):1432-6

Jamieson RV, Perveen R, Kerr B, Carette M, Yardley J, Heon E, Wirth MG, van H, V, Donnai D, Munier F, Black GC. (2002). Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum. Mol. Genet.* 11(1):33-42

Jefferson AB, Majerus PW. (1995). Properties of type II inositol polyphosphate 5-phosphatase. *J. Biol. Chem.* 270(16):9370-7

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302(5653):2141-4

Johnson K, Wodarz A. (2003). A genetic hierarchy controlling cell polarity. *Nat. Cell Biol.* 5(1):12-4

Jones SE, Jomary C, Grist J, Makwana J, Neal MJ. (1999). Retinal expression of gamma-crystallins in the mouse. *Invest Ophthalmol. Vis. Sci.* 40(12):3017-20

Jordan T, Hanson I, Zaletayev D, Hodgson S, Prosser J, Seawright A, Hastie N, van H, V. (1992). The human PAX6 gene is mutated in two patients with aniridia. *Nat. Genet.* 1(5):328-32

Juang JL, Hoffmann FM. (1999). Drosophila abelson interacting protein (dAbi) is a positive regulator of abelson tyrosine kinase activity. *Oncogene* 18(37):5138-47

Kannabiran C, Rogan PK, Olmos L, Basti S, Rao GN, Kaiser-Kupfer M, Hejtmancik JF. (1998). Autosomal dominant zonular cataract with sutural opacities is associated with a splice mutation in the betaA3/A1-crystallin gene. *Mol. Vis.* 4:21

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31(1):51-4

Katoh M, Katoh M. (2004). Identification and characterization of human GUKH2 gene in silico. *Int. J. Oncol.* 24(4):1033-8

Kelleher JF, Atkinson SJ, Pollard TD. (1995). Sequences, structural models, and cellular localization of the actin-related proteins Arp2 and Arp3 from Acanthamoeba. *J. Cell Biol.* 131(2):385-97

Kent OA, MacMillan AM. (2002). Early organization of pre-mRNA during spliceosome assembly. *Nat. Struct. Biol.* 9(8):576-81

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). The human genome browser at UCSC. *Genome Res.* 12(6):996-1006

Kessels MM, Engqvist-Goldstein AE, Drubin DG. (2000). Association of mouse actin-binding protein 1 (mAbp1/SH3P7), an Src kinase target, with dynamic regions of the cortical actin cytoskeleton in response to Rac1 activation. *Mol. Biol. Cell* 11(1):393-412

Kim AS, Kakalis LT, Abdul-Manan N, Liu GA, Rosen MK. (2000). Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature* 404(6774):151-8

Kobayashi K, Kuroda S, Fukata M, Nakamura T, Nagase T, Nomura N, Matsuura Y, Yoshida-Kubomura N, Iwamatsu A, Kaibuchi K. (1998). p140Sra-1 (specifically Rac1-associated protein) is a novel specific target for Rac1 small GTPase. *J. Biol. Chem.* 273(1):291-5

Koonin EV. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.*

Krill AE, Woodbury G, Bowman JE. (1969). X-chromosomal-linked sutural cataracts. *Am. J. Ophthalmol.* 68(5):867-72

Kruglyak L. (2005). Power tools for human genetics. *Nat. Genet.*

Kuszak JR, Bertram BA, Macsai MS, Rae JL. (1984). Sutures of the crystalline lens: a review. *Scan Electron Microsc.*(Pt 3):1369-78

Leahey AM, Charnas LR, Nussbaum RL. (1993). Nonsense mutations in the OCRL-1 gene in patients with the oculocerebrorenal syndrome of Lowe. *Hum. Mol. Genet.* 2(4):461-3

Lechler T, Shevchenko A, Li R. (2000). Direct involvement of yeast type I myosins in Cdc42-dependent actin polymerization. *J. Cell Biol.* 148(2):363-73

Lee WL, Bezanilla M, Pollard TD. (2000). Fission yeast myosin-I, Myo1p, stimulates actin assembly by Arp2/3 complex and shares functions with WASp. *J. Cell Biol.* 151(4):789-800

Lees-Miller JP, Henry G, Helfman DM. (1992). Identification of act2, an essential gene in the fission yeast Schizosaccharomyces pombe that encodes a protein related to actin. *Proc. Natl. Acad. Sci. U. S. A* 89(1):80-3

Leng Y, Zhang J, Badour K, Arpaia E, Freeman S, Cheung P, Siu M, Siminovitch K. (2005). Abelson-interactor-1 promotes WAVE2 membrane translocation and Abelson-mediated tyrosine phosphorylation required for WAVE2 activation. *Proc. Natl. Acad. Sci. U. S. A* 102(4):1098-103

Lewis RA. (1989). Mapping the gene for X-linked cataracts and microcornea with facial, dental, and skeletal features to Xp22: an appraisal of the Nance-Horan syndrome. *Trans. Am. Ophthalmol. Soc.* 87:658-728

Litt M, Carrero-Valenzuela R, LaMorticella DM, Schultz DW, Mitchell TN, Kramer P, Maumenee IH. (1997). Autosomal dominant cerulean cataract is associated with a chain

termination mutation in the human beta-crystallin gene CRYBB2. *Hum. Mol. Genet.* 6(5):665-8

Litt M, Kramer P, LaMorticella DM, Murphey W, Lovrien EW, Weleber RG. (1998). Autosomal dominant congenital cataract associated with a missense mutation in the human alpha crystallin gene CRYAA. *Hum. Mol. Genet.* 7(3):471-4

Lowe M. (2005). Structure and function of the Lowe syndrome protein OCRL1. *Traffic.* 6(9):711-9

Lubsen NH, Aarts HJ, Schoenmakers JG. (1988). The evolution of lenticular proteins: the beta- and gamma-crystallin super gene family. *Prog. Biophys. Mol. Biol.* 51(1):47-76

Machesky LM, Atkinson SJ, Ampe C, Vandekerckhove J, Pollard TD. (1994). Purification of a cortical complex containing two unconventional actins from Acanthamoeba by affinity chromatography on profilin-agarose. *J. Cell Biol.* 127(1):107-15

Machesky LM, Reeves E, Wientjes F, Mattheyse FJ, Grogan A, Totty NF, Burlingame AL, Hsuan JJ, Segal AW. (1997). Mammalian actin-related protein 2/3 complex localizes to regions of lamellipodial protrusion and is composed of evolutionarily conserved proteins. *Biochem. J.* 328 (Pt 1):105-12

Machesky LM, Insall RH. (1998). Scar1 and the related Wiskott-Aldrich syndrome protein, WASP, regulate the actin cytoskeleton through the Arp2/3 complex. *Curr. Biol.* 8(25):1347-56

Machesky LM, Gould KL. (1999). The Arp2/3 complex: a multifunctional actin organizer. *Curr. Opin. Cell Biol.* 11(1):117-21

Machesky LM, Mullins RD, Higgs HN, Kaiser DA, Blanchoin L, May RC, Hall ME, Pollard TD. (1999). Scar, a WASp-related protein, activates nucleation of actin filaments by the Arp2/3 complex. *Proc. Natl. Acad. Sci. U. S. A* 96(7):3739-44

Mackay D, Ionides A, Kibar Z, Rouleau G, Berry V, Moore A, Shiels A, Bhattacharya S. (1999). Connexin46 mutations in autosomal dominant congenital cataract. *Am. J. Hum. Genet.* 64(5):1357-64

Mackay DS, Boskovska OB, Knopf HL, Lampi KJ, Shiels A. (2002). A nonsense mutation in CRYBB1 associated with autosomal dominant cataract linked to human chromosome 22q. *Am. J. Hum. Genet.* 71(5):1216-21

Mackay DS, Andley UP, Shiels A. (2004). A missense mutation in the gammaD crystallin gene (CRYGD) associated with autosomal dominant "coral-like" cataract linked to chromosome 2q. *Mol. Vis.* 10:155-62

Maisel H, Perry MM. (1972) Electron microscope observations on some structural proteins of the chick lens. *Exp Eye Res.*14(1):7-12.

Marchand JB, Kaiser DA, Pollard TD, Higgs HN. (2001). Interaction of WASP/Scar proteins with actin and vertebrate Arp2/3 complex. *Nat. Cell Biol.* 3(1):76-82

Martha A, Strong LC, Ferrell RE, Saunders GF. (1995). Three novel aniridia mutations in the human PAX6 gene. *Hum. Mutat.* 6(1):44-9

Mathew S, Chaudhuri A, Murty VV, Pogo AO. (1994). Confirmation of Duffy blood group antigen locus (FY) at 1q22-->q23 by fluorescence in situ hybridization. *Cytogenet. Cell Genet.* 67(1):68

Mathias RT, Rae JL, Baldo GJ. (1997). Physiological properties of the normal lens. *Physiol Rev.* 77(1):21-50

McAvoy J. (1981). In *Developmental Biology of the Lens in Mechanisms*,7-46 pp. Academic Press. 7-46 pp.

McCollum D, Feoktistova A, Morphew M, Balasubramanian M, Gould KL. (1996). The Schizosaccharomyces pombe actin-related protein, Arp3, is a component of the cortical actin cytoskeleton and interacts with profilin. *EMBO J.* 15(23):6438-46

McConkey EH. (2004). Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet. Genome Res.* 105(1):157-8

McKay JD, Patterson B, Craig JE, Russell-Eggitt IM, Wirth MG, Burdon KP, Hewitt AW, Cohn AC, Kerdraon Y, Mackey DA. (2005). The telomere of human chromosome 1p contains at least two independent autosomal dominant congenital cataract genes. *Br. J. Ophthalmol.* 89(7):831-4

Miki H, Suetsugu S, Takenawa T. (1998). WAVE, a novel WASP-family protein involved in actin reorganization induced by Rac. *EMBO J.* 17(23):6932-41

Miki H, Takenawa T. (1998). Direct binding of the verprolin-homology domain in N-WASP to actin is essential for cytoskeletal reorganization. *Biochem. Biophys. Res. Commun.* 243(1):73-8

Mitchison TJ, Cramer LP. (1996). Actin-based cell motility and cell locomotion. *Cell* 84(3):371-9

Miyazaki K, Matsuda S, Ichigotani Y, Takenouchi Y, Hayashi K, Fukuda Y, Nimura Y, Hamaguchi M. (2000). Isolation and characterization of a novel human gene (NESH) which encodes a putative signaling molecule similar to e3B1 protein. *Biochim. Biophys. Acta* 1493(1-2):237-41

Mullins RD, Heuser JA, Pollard TD. (1998). The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proc. Natl. Acad. Sci. U. S. A* 95(11):6181-6

Neuhaus JM, Wanger M, Keiser T, Wegner A. (1983). Treadmilling of actin. *J. Muscle Res. Cell Motil.* 4(5):507-27

Ng D, Thakker N, Corcoran CM, Donnai D, Perveen R, Schneider A, Hadley DW, Tifft C, Zhang L, Wilkie AO, van der Smagt JJ, Gorlin RJ, Burgess SM, Bardwell VJ, Black GC, Biesecker LG. (2004). Oculofaciocardiodental and Lenz microphthalmia syndromes result from distinct classes of mutations in BCOR. *Nat. Genet.* 36(4):411-6

Nogami S, Satoh S, Nakano M, Shimizu H, Fukushima H, Maruyama A, Terano A, Shirataki H. (2003). Taxilin; a novel syntaxin-binding protein that is involved in Ca2+-dependent exocytosis in neuroendocrine cells. *Genes Cells* 8(1):17-28

Nussbaum RL, Orrison BM, Janne PA, Charnas L, Chinault AC. (1997). Physical mapping and genomic structure of the Lowe syndrome gene OCRL1. *Hum. Genet.* 99(2):145-50

Ogino H, Yasuda K. (1998). Induction of lens differentiation by activation of a bZIP transcription factor, L-Maf. *Science* 280(5360):115-8

Oikawa T, Yamaguchi H, Itoh T, Kato M, Ijuin T, Yamazaki D, Suetsugu S, Takenawa T. (2004). PtdIns(3,4,5)P3 binding is necessary for WAVE2-induced formation of lamellipodia. *Nat. Cell Biol.* 6(5):420-6

Panchal SC, Kaiser DA, Torres E, Pollard TD, Rosen MK. (2003). A conserved amphipathic helix in WASP/Scar protein is essential for activation of Arp2/3 complex. *Nat Struct Biol.*10(8):591-8.

Parthun MR, Widom J, Gottschling DE. (1996). The major cytoplasmic histone acetyltransferase in yeast: links to chromatin replication and histone metabolism. *Cell* 87(1):85-94

Paznekas WA, Boyadjiev SA, Shapiro RE, Daniels O, Wollnik B, Keegan CE, Innis JW, Dinulos MB, Christian C, Hannibal MC, Jabs EW. (2003). Connexin 43 (GJA1) mutations cause the pleiotropic phenotype of oculodentodigital dysplasia. *Am. J. Hum. Genet.* 72(2):408-18

Pennisi E. (2003). Human genome. Reaching their goal early, sequencing labs celebrate. *Science* 300(5618):409

Peterson LC, Rao KV, Crosson JT, White JG. (1985) Fechtner syndrome-a variant of Alport's syndrome with leukocyte inclusions and macrothrombocytopenia. *Blood.* 65(2):397-406.

Pinna A, Carta A, Mannazzu MC, Dore S, Balata A, Carta F. (2004). WAGR syndrome with deletion of chromosome 11p11.2-13. *J. AAPOS.* 8(4):396-7

Pollard TD, Blanchoin L, Mullins RD. (2000). Molecular mechanisms controlling actin filament dynamics in nonmuscle cells. *Annu. Rev. Biophys. Biomol. Struct.* 29:545-76

Pras E, Frydman M, Levy-Nissenbaum E, Bakhan T, Raz J, Assia EI, Goldman B, Pras E. (2000). A nonsense mutation (W9X) in CRYAA causes autosomal recessive cataract in an inbred Jewish Persian family. *Invest Ophthalmol. Vis. Sci.* 41(11):3511-5

Pras E, Levy-Nissenbaum E, Bakhan T, Lahat H, Assia E, Geffen-Carmi N, Frydman M, Goldman B, Pras E. (2002). A missense mutation in the LIM2 gene is associated with autosomal recessive presenile cataract in an inbred Iraqi Jewish family. *Am. J. Hum. Genet.* 70(5):1363-7

Pras E, Raz J, Yahalom V, Frydman M, Garzozi HJ, Pras E, Hejtmancik JF. (2004). A nonsense mutation in the glucosaminyl (N-acetyl) transferase 2 gene (GCNT2): association with autosomal recessive congenital cataracts. *Invest Ophthalmol. Vis. Sci.* 45(6):1940-5

Prehoda KE, Scott JA, Mullins RD, Lim WA. (2000). Integration of multiple signals through cooperative regulation of the N-WASP-Arp2/3 complex. *Science* 290(5492):801-6

Qian YW, Lee EY. (1995). Dual retinoblastoma-binding proteins with properties related to a negative regulator of ras in yeast. *J. Biol. Chem.* 270(43):25507-13

Ramprasad VL, Thool A, Murugan S, Nancarrow D, Vyas P, Rao SK, Vidhya A, Ravishankar K, Kumaramanickavel G. (2005). Truncating mutation in the NHS gene: phenotypic heterogeneity of Nance-Horan syndrome in an asian Indian family. *Invest Ophthalmol. Vis. Sci.* 46(1):17-23

Raucher D, Stauffer T, Chen W, Shen K, Guo S, York JD, Sheetz MP, Meyer T. (2000). Phosphatidylinositol 4,5-bisphosphate functions as a second messenger that regulates cytoskeleton-plasma membrane adhesion. *Cell* 100(2):221-8

Reaume AG, de Sousa PA, Kulkarni S, Langille BL, Zhu D, Davies TC, Juneja SC, Kidder GM, Rossant J. (1995). Cardiac malformation in neonatal mice lacking connexin43. *Science* 267(5205):1831-4

Reddy BA, Kloc M, Etkin LD. (1992). The cloning and characterization of a localized maternal transcript in Xenopus laevis whose zygotic counterpart is detected in the CNS. *Mech. Dev.* 39(3):143-50

Reddy MA, Francis PJ, Berry V, Bhattacharya SS, Moore AT. (2004). Molecular genetic basis of inherited cataract and associated phenotypes. *Surv. Ophthalmol.* 49(3):300-15

Reilly DS, Lewis RA, Nussbaum RL. (1990). Genetic and physical mapping of Xq24-q26 markers flanking the Lowe oculocerebrorenal syndrome. *Genomics* 8(1):62-70

Ren Z, Li A, Shastry BS, Padma T, Ayyagari R, Scott MH, Parks MM, Kaiser-Kupfer MI, Hejtmancik JF. (2000). A 5-base insertion in the gammaC-crystallin gene is associated with autosomal dominant variable zonular pulverulent cataract. *Hum Genet.* 106(5):531-7

Renwick JH, Lawler SD. (1963). Probable linkage between a congenital cataract locus and the duffy blood group locus. *Ann. Hum. Genet.* 27:67-84

Riazuddin SA, Yasmeen A, Yao W, Sergeev YV, Zhang Q, Zulfiqar F, Riaz A, Riazuddin S, Hejtmancik JF. (2005). Mutations in betaB3-crystallin associated with autosomal recessive cataract in two Pakistani families. *Invest Ophthalmol. Vis. Sci.* 46(6):2100-6

Riazuddin SA, Yasmeen A, Zhang Q, Yao W, Sabar MF, Ahmed Z, Riazuddin S, Hejtmancik JF. (2005)b. A new locus for autosomal recessive nuclear cataract mapped to chromosome 19q13 in a Pakistani family. *Invest Ophthalmol. Vis. Sci.* 46(2):623-6

Rohatgi R, Ma L, Miki H, Lopez M, Kirchhausen T, Takenawa T, Kirschner MW. (1999) The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. *Cell.* 97(2):221-31.

Rohatgi R, Ho HY, Kirschner MW. (2000). Mechanism of N-WASP activation by CDC42 and phosphatidylinositol 4, 5-bisphosphate. *J. Cell Biol.* 150(6):1299-310

Rohatgi R, Nollau P, Ho HY, Kirschner MW, Mayer BJ. (2001). Nck and phosphatidylinositol 4,5-bisphosphate synergistically activate actin polymerization through the N-WASP-Arp2/3 pathway. *J. Biol. Chem.* 276(28):26448-52

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239(4839):487-91

Schmid AC, Wise HM, Mitchell CA, Nussbaum R, Woscholski R. (2004). Type II phosphoinositide 5-phosphatases have unique sensitivities towards fatty acid composition and head group phosphorylation. *FEBS Lett.* 576(1-2):9-13

Schwob E, Martin RP. (1992). New yeast actin-like gene required late in the cell cycle. *Nature* 355(6356):179-82

Scita G, Nordstrom J, Carbone R, Tenca P, Giardina G, Gutkind S, Bjarnegard M, Betsholtz C, Di Fiore PP. (1999). EPS8 and E3B1 transduce signals from Ras to Rac. *Nature* 401(6750):290-3

Scita G, Tenca P, Areces LB, Tocchetti A, Frittoli E, Giardina G, Ponzanelli I, Sini P, Innocenti M, Di Fiore PP. (2001). An effector region in Eps8 is responsible for the activation of the Rac-specific GEF activity of Sos-1 and for the proper localization of the Rac-based actin-polymerizing machine. *J. Cell Biol.* 154(5):1031-44

Semina EV, Ferrell RE, Mintz-Hittner HA, Bitoun P, Alward WL, Reiter RS, Funkhauser C, Daack-Hirsch S, Murray JC. (1998). A novel homeobox gene PITX3 is mutated in families with autosomal-dominant cataracts and ASMD. *Nat. Genet.* 19(2):167-70

Semina EV, Murray JC, Reiter R, Hrstka RF, Graw J.(2000) Deletion in the promoter region and altered expression of Pitx3 homeobox gene in aphakia mice. *Hum Mol Genet.*9(11):1575-85.

Seow WK, Brown JP, Romaniuk K. (1985). The Nance-Horan syndrome of dental anomalies, congenital cataracts, microphthalmia, and anteverted pinna: case report. *Pediatr. Dent.* 7(4):307-11

Seri M, Cusano R, Gangarossa S, Caridi G, Bordo D, Lo NC, Ghiggeri GM, Ravazzolo R, Savino M, Del Vecchio M, d'Apolito M, Iolascon A, Zelante LL, Savoia A, Balduini CL, Noris P, Magrini U, Belletti S, Heath KE, Babcock M, Glucksman MJ, Aliprandis E, Bizzaro N, Desnick RJ, Martignetti JA. (2000). Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. The May-Heggllin/Fechtner Syndrome Consortium. *Nat. Genet.* 26(1):103-5

Shi Y, Alin K, Goff SP. (1995). Abl-interactor-1, a novel SH3 protein binding to the carboxy-terminal portion of the Abl protein, suppresses v-abl transforming activity. *Genes Dev.* 9(21):2583-97

Shiels A, Griffin CS. (1993). Aberrant expression of the gene for lens major intrinsic protein in the CAT mouse. *Curr. Eye Res.* 12(10):913-21

Shiels A, Bassnett S. (1996). Mutations in the founder of the MIP gene family underlie cataract development in the mouse. *Nat. Genet.* 12(2):212-5

Shiels A, Mackay D, Ionides A, Berry V, Moore A, Bhattacharya S. (1998). A missense mutation in the human connexin50 gene (GJA8) underlies autosomal dominant "zonular pulverulent" cataract, on chromosome 1q. *Am. J. Hum. Genet.* 62(3):526-32

Shimada N, Aya-Murata T, Reza HM, Yasuda K. (2003). Cooperative action between L-Maf and Sox2 on delta-crystallin gene expression during chick lens development. *Mech. Dev.* 120(4):455-65

Silver DN, Lewis RA, Nussbaum RL. (1987). Mapping the Lowe oculocerebrorenal syndrome to Xq24-q26 by use of restriction fragment length polymorphisms. *J. Clin. Invest* 79(1):282-5

Small JV, Rohlfs A, Herzog M. (1993). Actin and cell movement. *Symp Soc Exp Biol.*47:57-71.

Small JV. (1994). Lamellipodia architecture: actin filament turnover and the lateral flow of actin filaments during motility. *Semin. Cell Biol.* 5(3):157-63

Small JV. (1995). Getting the actin filaments straight: nucleation-release or treadmilling? *Trends Cell Biol.* 5(2):52-5

Small JV, Rottner K, Kaverina I, Anderson KI. (1998). Assembling an actin cytoskeleton for cell attachment and movement. *Biochim. Biophys. Acta* 1404(3):271-81

Small JV, Stradal T, Vignal E, Rottner K. (2002). The lamellipodium: where motility begins. *Trends Cell Biol.* 12(3):112-20

Smaoui N, Beltaief O, BenHamed S, M'Rad R, Maazoul F, Ouertani A, Chaabouni H, Hejtmancik JF. (2004). A homozygous splice mutation in the HSF4 gene is associated with an autosomal recessive congenital cataract. *Invest Ophthalmol. Vis. Sci.* 45(8):2716-21

Snell R, Lemp M. (1989). In *Clinical Anatomy of the Eye*, Blackwell Scientific.

Soderling SH, Binns KL, Wayman GA, Davee SM, Ong SH, Pawson T, Scott JD. (2002) The WRP component of the WAVE-1 complex attenuates Rac-mediated signalling. *Nat Cell Biol.* 4(12):970-5.

Soderling SH, Langeberg LK, Soderling JA, Davee SM, Simerly R, Raber J, Scott JD. (2003). Loss of WAVE-1 causes sensorimotor retardation and reduced learning and memory in mice. *Proc. Natl. Acad. Sci. U. S. A* 100(4):1723-8

Sofaer JA (1969). Aspects of the tabby-crinkled-downless syndrome. II. Observations on the reaction to changes of genetic background. *J Embryol Exp Morphol.* 22(2):207-27.

Sofaer JA (1977). The teeth of the "sleek" mouse. *Arch Oral Biol.* 22(4):299-301

Stambolian D, Lewis RA, Buetow K, Bond A, Nussbaum R. (1990). Nance-Horan syndrome: localization within the region Xp21.1-Xp22.3 by linkage analysis. *Am. J. Hum. Genet.* 47(1):13-9

Stambolian D, Favor J, Silvers W, Avner P, Chapman V, Zhou E. (1994). Mapping of the X-linked cataract (Xcat) mutation, the gene implicated in the Nance Horan syndrome, on the mouse X chromosome. *Genomics* 22(2):377-80

Steele EC, Jr., Kerscher S, Lyon MF, Glenister PH, Favor J, Wang J, Church RL. (1997). Identification of a mutation in the MP19 gene, Lim2, in the cataractous mouse mutant To3. *Mol. Vis.* 3:5

Stephan DA, Gillanders E, Vanderveen D, Freas-Lutz D, Wistow G, Baxevanis AD, Robbins CM, VanAuken A, Quesenberry MI, Bailey-Wilson J, Juo SH, Trent JM, Smith L, Brownstein MJ. (1999). Progressive juvenile-onset punctate cataracts caused by mutation of the gammaD-crystallin gene. *Proc. Natl. Acad. Sci. U. S. A* 96(3):1008-12

Stradal T, Courtney KD, Rottner K, Hahne P, Small JV, Pendergast AM. (2001). The Abl interactor proteins localize to sites of actin polymerization at the tips of lamellipodia and filopodia. *Curr. Biol.* 11(11):891-5

Stradal TE, Rottner K, Disanza A, Confalonieri S, Innocenti M, Scita G. (2004). Regulation of actin dynamics by WASP and WAVE family proteins. *Trends Cell Biol.* 14(6):303-11

Suchy SF, Olivos-Glander IM, Nussabaum RL. (1995). Lowe syndrome, a deficiency of phosphatidylinositol 4,5-bisphosphate 5-phosphatase in the Golgi apparatus. *Hum. Mol. Genet.* 4(12):2245-50

Suchy SF, Nussbaum RL. (2002). The deficiency of PIP2 5-phosphatase in Lowe syndrome affects actin polymerization. *Am. J. Hum. Genet.* 71(6):1420-7

Suetsugu S, Miki H, Takenawa T. (1999). Identification of two human WAVE/SCAR homologues as general actin regulatory molecules which associate with the Arp2/3 complex. *Biochem. Biophys. Res. Commun.* 260(1):296-302

Suetsugu S, Hattori M, Miki H, Tezuka T, Yamamoto T, Mikoshiba K, Takenawa T. (2002). Sustained activation of N-WASP through phosphorylation is essential for neurite extension. *Dev. Cell* 3(5):645-58

Sun H, Ma Z, Li Y, Liu B, Li Z, Ding X, Gao Y, Ma W, Tang X, Li X, Shen Y. (2005). Gamma-S crystallin gene (CRYGS) mutation causes dominant progressive cortical cataract in humans. *J. Med. Genet.* 42(9):706-10

Svitkina TM, Borisy GG. (1999). Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia. *J. Cell Biol.* 145(5):1009-26

Symons M, Derry JM, Karlak B, Jiang S, Lemahieu V, Mccormick F, Francke U, Abo A. (1996). Wiskott-Aldrich syndrome protein, a novel effector for the GTPase CDC42Hs, is implicated in actin polymerization. *Cell* 84(5):723-34

Szpiro-Tapia S, Bobrie G, Guilloud-Bataille M, Heuertz S, Julier C, Frezal J, Grunfeld JP, Hors-Cayla MC. (1988). Linkage studies in X-linked Alport's syndrome. *Hum. Genet.* 81(1):85-7

Takenawa T, Miki H. (2001). WASP and WAVE family proteins: key molecules for rapid rearrangement of cortical actin filaments and cell movement. *J. Cell Sci.* 114(Pt 10):1801-9

Tanentzapf G, Tepass U. (2003). Interactions between the crumbs, lethal giant larvae and bazooka pathways in epithelial polarization. *Nat. Cell Biol.* 5(1):46-52

ter Haar E, Harrison SC, Kirchhausen T. (2000). Peptide-in-groove interactions link target proteins to the beta-propeller of clathrin. *Proc. Natl. Acad. Sci. U. S. A* 97(3):1096-100

The human genome consortium (2001). The Human Genome. *Nature* 409:745-964

Toker A. (1998). The synthesis and cellular roles of phosphatidylinositol 4,5-bisphosphate. *Curr. Opin. Cell Biol.* 10(2):254-61

Toutain A, Ayrault AD, Moraine C. (1997). Mental retardation in Nance-Horan syndrome: clinical and neuropsychological assessment in four families. *Am. J. Med. Genet.* 71(3):305-14

Toutain A, Ronce N, Dessay B, Robb L, Francannet C, Le Merrer M, Briard ML, Kaplan J, Moraine C. (1997). Nance-Horan syndrome: linkage analysis in 4 families refines localization in Xp22.31-p22.13 region. *Hum. Genet.* 99(2):256-61

Toutain A, Dessay B, Ronce N, Ferrante MI, Tranchemontagne J, Newbury-Ecob R, Wallgren-Pettersson C, Burn J, Kaplan J, Rossi A, Russo S, Walpole I, Hartsfield JK, Oyen N, Nemeth A, Bitoun P, Trump D, Moraine C, Franco B. (2002). Refinement of the NHS locus on chromosome Xp22.13 and analysis of five candidate genes. *Eur. J. Hum. Genet.* 10(9):516-20

Tucker A and Sharpe P. (2004). The cutting-edge of mammalian development; how the embryo makes teeth. *Nat Rev Genet.* 2004. 5(7):499-508.

Ungewickell A, Ward ME, Ungewickell E, Majerus PW. (2004). The inositol polyphosphate 5-phosphatase Ocrl associates with endosomes that are partially coated with clathrin. *Proc. Natl. Acad. Sci. U. S. A* 101(37):13501-6

Uruno T, Liu J, Zhang P, Fan Yx, Egile C, Li R, Mueller SC, Zhan X. (2001). Activation of Arp2/3 complex-mediated actin polymerization by cortactin. *Nat Cell Biol.* 3(3):259-66.

van Reeuwijk J, Janssen M, van den EC, Beltran-Valero dB, Sabatelli P, Merlini L, Boon M, Scheffer H, Brockington M, Muntoni F, Huynen M, Verrips A, Walsh C, Barth P, Brunner H, van Bokhoven H. (2005). POMT2 mutations cause alpha-dystroglycan hypoglycosylation and Walker Warburg syndrome. *J. Med. Genet.*

Vasioukhin V, Bauer C, Yin M, Fuchs E. (2000). Directed actin polymerization is the driving force for epithelial cell-cell adhesion. *Cell* 100(2):209-19

Vermaak D, Wade PA, Jones PL, Shi YB, Wolffe AP. (1999). Functional analysis of the SIN3-histone deacetylase RPD3-RbAp48-histone H4 connection in the Xenopus oocyte. *Mol. Cell Biol.* 19(9):5847-60

Verreault A, Kaufman PD, Kobayashi R, Stillman B. (1998). Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase. *Curr. Biol.* 8(2):96-108

Walpole SM, Ronce N, Grayson C, Dessay B, Yates JR, Trump D, Toutain A. (1999). Exclusion of RAI2 as the causative gene for Nance-Horan syndrome. *Hum. Genet.* 104(5):410-1

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES, . (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280(5366):1077-82

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade

CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520-62

Weaver AM, Karginov AV, Kinley AW, Weed SA, Li Y, Parsons JT, Cooper JA. (2001). Cortactin promotes and stabilizes Arp2/3-induced actin filament network formation. *Curr. Biol.* 11(5):370-4

Weaver AM, Heuser JE, Karginov AV, Lee WL, Parsons JT, Cooper JA. (2002). Interaction of cortactin and N-WASp with Arp2/3 complex. *Curr. Biol.* 12(15):1270-8

Weed SA, Karginov AV, Schafer DA, Weaver AM, Kinley AW, Cooper JA, Parsons JT. (2000). Cortactin localization to sites of actin assembly in lamellipodia requires interactions with F-actin and the Arp2/3 complex. *J. Cell Biol.* 151(1):29-40

Wegner A. (1976). Head to tail polymerization of actin. *J. Mol. Biol.* 108(1):139-50

Welch MD, DePace AH, Verma S, Iwamatsu A, Mitchison TJ. (1997). The human Arp2/3 complex is composed of evolutionarily conserved subunits and is localized to cellular regions of dynamic actin filament assembly. *J. Cell Biol.* 138(2):375-84

Welch MD, Rosenblatt J, Skoble J, Portnoy DA, Mitchison TJ. (1998). Interaction of human Arp2/3 complex and the Listeria monocytogenes ActA protein in actin filament nucleation. *Science* 281(5373):105-8

Wenk MR, Lucast L, Di Paolo G, Romanelli AJ, Suchy SF, Nussbaum RL, Cline GW, Shulman GI, McMurray W, De Camilli P. (2003). Phosphoinositide profiling in complex lipid mixtures using electrospray ionization mass spectrometry. *Nat. Biotechnol.* 21(7):813-7

White TW, Bruzzone R, Goodenough DA, Paul DL. (1992). Mouse Cx50, a functional member of the connexin family of gap junction proteins, is the lens fiber protein MP70. *Mol. Biol. Cell* 3(7):711-20

White TW, Goodenough DA, Paul DL. (1998). Targeted ablation of connexin50 in mice results in microphthalmia and zonular pulverulent cataracts. *J. Cell Biol.* 143(3):815-25

Wilkie AO, Taylor D, Scambler PJ, Baraitser M. (1993). Congenital cataract, microphthalmia and septal heart defect in two generations: a new syndrome? *Clin. Dysmorphol.* 2(2):114-9

Winter D, Podtelejnikov AV, Mann M, Li R. (1997). The complex containing actin-related proteins Arp2 and Arp3 is required for the motility and integrity of yeast actin patches. *Curr. Biol.* 7(7):519-29

Winter D, Lechler T, Li R. (1999). Activation of the yeast Arp2/3 complex by Bee1p, a WASP-family protein. *Curr. Biol.* 9(9):501-4

Winter DC, Choe EY, Li R. (1999). Genetic dissection of the budding yeast Arp2/3 complex: a comparison of the in vivo and structural roles of individual subunits. *Proc. Natl. Acad. Sci. U. S. A* 96(13):7288-93

Wolfsberg TG, Madden TL. (1999). *Current Protocols in Molecular Biology*:19.31-19.3.29

Wolfsberg TG, McEntyre J, Schuler GD. (2001). Guide to the draft human genome. *Nature* 409(6822):824-6

Yamaguchi H, Miki H, Takenawa T. (2002) Two verprolin homology domains increase the Arp2/3 complex-mediated actin polymerization activities of N-WASP and WAVE1 C-terminal regions. *Biochem Biophys Res Commun.* 297(2):214-9.

Yang J, Kiefer S, Rauchman M. (2002). Characterization of the gene encoding mouse retinoblastoma binding protein-7, a component of chromatin-remodeling complexes. *Genomics* 80(4):407-15

Yarar D, To W, Abo A, Welch MD. (1999). The Wiskott-Aldrich syndrome protein directs actin-based motility by stimulating actin nucleation with the Arp2/3 complex. *Curr. Biol.* 9(10):555-8

Yarar D, To W, Abo A, Welch MD. (1999). The Wiskott-Aldrich syndrome protein directs actin-based motility by stimulating actin nucleation with the Arp2/3 complex. *Curr. Biol.* 9(10):555-8

Zalevsky J, Grigorova I, Mullins RD. (2001). Activation of the Arp2/3 complex by the Listeria acta protein. Acta binds two actin monomers and three subunits of the Arp2/3 complex. *J. Biol. Chem.* 276(5):3468-75

Zallen JA, Cohen Y, Hudson AM, Cooley L, Wieschaus E, Schejter ED. (2002). SCAR is a primary regulator of Arp2/3-dependent morphological events in Drosophila. *J. Cell Biol.* 156(4):689-701

Zhang X, Jefferson AB, Auethavekiat V, Majerus PW. (1995). The protein deficient in Lowe syndrome is a phosphatidylinositol-4,5-bisphosphate 5-phosphatase. *Proc. Natl. Acad. Sci. U. S. A* 92(11):4853-6

Zhang X, Hartz PA, Philip E, Racusen LC, Majerus PW. (1998). Cell lines from kidney proximal tubules of a patient with Lowe syndrome lack OCRL inositol polyphosphate 5-phosphatase and accumulate phosphatidylinositol 4,5-bisphosphate. *J. Biol. Chem.* 273(3):1574-82

Zhang Q, Guo X, Xiao X, Yi J, Jia X, Hejtmancik JF. (2004). Clinical description and genome wide linkage study of Y-sutural cataract and myopia in a Chinese family. *Mol. Vis.* 10:890-900

Zhou L, Chen T, Church RL. (2002). Temporal expression of three mouse lens fiber cell membrane protein genes during early development. *Mol. Vis.* 8:143-8

Zhu D, Alcorn DM, Antonarakis SE, Levin LS, Huang PC, Mitchell TN, Warren AC, Maumenee IH. (1990). Assignment of the Nance-Horan syndrome to the distal short arm of the X chromosome. *Hum. Genet.* 86(1):54-8

Ziemnicka-Kotula D, Xu J, Gu H, Potempska A, Kim KS, Jenkins EC, Trenkner E, Kotula L. (1998). Identification of a candidate human spectrin Src homology 3 domain-binding protein suggests a general mechanism of association of tyrosine kinases with the spectrin-based membrane skeleton. *J. Biol. Chem.* 273(22):13681-92

- Appendix A -

Protein sequences of the NHS paralogs (NHSL1, NHSL2, and NHSL3)

Protein sequences of the NHS paralogs are shown below. Putative bipartite nuclear localisation signals identified in NHSL1 and NHSL2 are highlighted in red and underlined.

NHSL1

MPFHQRTLEPARLRRPEAAGAGAAGAPLFRSLEQVSSHALGCLLAQLADLSRCAGDIFGELEGQ
AAALGHRTAALHRRLDALQAAAARLDHRRVKIPVSNLDEESRWTVHYTAPWHQQENVFLPTTRP
PCVEDLHRQAKLNLKSVLRECDKLRHDGYRSSQYYSQGPTFAANASPFCDDYQDEDEETDQKCS
LSSSEEERFISIRRPKTPASSDFSDLNTQTNWTKSLPLPTPEEKMRQQAQTVQADVVPINITGE
NFDRQASLRRSLIYTDTLVRRPKKVKRRKTITGVPDNIQKELASGTGQDDADGHSVYTPDHYST
LGRFNSCRSAGQRSETRDSSCQTEDVKVVPPSMRRIRAQKGQGIAAQMGHFSGSSGNMSVLSDS
AGIVFPSRLDSDAGFHSLPRSGARANIQSLEPRLGALGPAGDMNGTFLYQRGHPQADENLGHLG
GASGTGTLLRPKSQELRHFESENIMSPACVVSPHATYSTSIIPNATLSSSSEVIAIPTAQSAGQ
RESKSSGSSHARIKSRDHLISRHAVKGDPQSPGRHWNEGHATILSQDLDPHSPGEPALLSLCDS
AVPLNAPANRENGSQAMPYNCRNNLAFPAHPQDVDGKSESSYSGGGGHSSSEPWEYKSSGNGRA
SPLKPHLATPGYSTPTSNMSSCSLDQTSNKEDAGSLYSEDHDGYCASVHTDSGHGSGNLCNSSD
GFGNPRHSVINVFVGRAQKNQGDRSNYQDKSLSRNISLKKAKKPPLPPSRTDSLRRIPKKSSQC
NGQVLNESLIATLQHSLQLSLPGKSGSSPSQSPCSDLEEPWLPRSRSQSTVSAGSSMTSATTPN
VYSLCGATPSQSDTSSVKSEYTDPWGYYIDYTGMQEDPGNPAGGCSTSSGVPTGNGPVRHVQEG
SRATMPQVPGGSVKPKIMSPEKSHRVISPSSGYSSQSNTPTALTPVPVFLKSVSPANGKGKPKP
KVPERKSSLISSVSISSSSTSLSSSTSTEGSGTMKKLDPAVGSPPAPPPPPVPSPPFPCPADRS
PFLPPPPPVTDCSQGSPLPHSPVFPPPPPEALIPFCSPPDWCLSPPRPALSPILPDSPVSLPLP
PPLLPSSEPPPAPPLDPKFMKDTRPPFTNSGQPESSRGSLRPPSTKEETSRPPMPLITTEALQM
VQLRPVRKNSGAEAAQLSERTAQEQRTPVAPQYHLKPSAFLKSRNSTNEMESESQPASVTSSLP
TPAKSSSQGDHGSAAERGGPVSRSPGAPSAGEAEARPSPSTTPLPDSSPSRKPPPISKKPKLFL
VVPPPQKDFAVEPAENVSEALRAVPSPTTGEEGSVHSREAKESSAAQAGSHATHPGTSVLEGGA
AGSMSPSRVEANVPMVQPDVSPAPKQEEPAENSADTGGDGESCLSQQDGAAGVPETNAAGSSSE
ACDFLKEDGNDEVMTPSRPRTTEDLFAAIHRSKRKVLGRRDSDDDHSRNHSPSPPVTPTGAAPS
LASPKQVGSIQRSIRKSSTSSDNFKALLLKKGSRSDTSARMSAAEMLKNTDPRFQRSRSEPSPD
APESPSSCSPSKNRRAQEEWAKNEGLMPRSLSFSGPRYGRSRTPPSAASSRYSMRNRIQSSPMT
VISEGEGEAVEPVDSIARGALGAAEGCSLDGLAREEMDEGGLLCGEGPAASLQPQAPGPVDGTA
SAEGREPSPQCGGSLSEES

NHSL2

MPFYRRTVVPQRLCPRNPPQQLAELRDVSHLAALSLLRQLADLCGHSLALLEDLEGHLLALGRR
TDSLYRRTVRLRRRLPCRLLGPEEDEEELGAANSGRENATATAHSRSSWRQPVNVFLSSGRPPS
VEELLREAQLNLQSLLQEEYEEQYSEARLVGQTFRSSDEATKPTPNPRPQSARRLEFILMPTKR
QLSEDETTTQGVRAPEASLSLSTTADKQTAWNSLFPLPILEEKRWPQLCSTQSDIVPINISETR
MHPSVFILATGQQFDKHASLRHSLFNTETAVNPKSTLRRRRTIIGFSNFSQRDQGHSNSPAGSV
AHSTTSDIRPSHSVPEGVHGRVAVGQDARFPSLTSPVLRTPSSEPDEPHQARSGPNPPGMESMG
MVYSVPSSCNGPTESTFSTSWKGDAFTYMTPSATSQSNQVNENGKNPSCGNSWVSLNKVPPLVP
KEAATLLVARDNPAGCSGSAGYPERLIQQRHMPERPSKIGLLTSGTSRLETGPGGASRFRERSL
SVPTDSGTTDVDYDEEQKANEACALPFASTSSEGSNSADNIASLSAQQEAQHRRQRSKSISLRK
AKKKPSPPTRSVSLVKDEPGLLPEGGSALPKDQRPKSLCLSLEHQGHHSSHPDAQGHPAIPNHK
DPESTQFSHHWYLTDWKSGDTYQSLSSSSTATGTTVIECTQVQGSSESLASPSTSRATTPSQLS
IEVEAREISSPGRPPGLMSPSSGYSSQSETPTPTVSMSLTLGHLPPPSSSVRVRPVVPERKSSL
PPTSPMEKFPKSRLSFDLPLTSSPNLDLSGMSISIRSKTKVSRHHSETNFGVKLAQKTNPNQPI
MPMVTQSDLRSVRLRSVSKSEPEDDIESPEYAEEPRAEEVFTLPERKTKPPVAEKPPVARRPPS
LVHKPPSVPEEYALTSPTLAMPPRSSIQHARPLPQDSYTVVRKPKPSSFPDGRSPGESTAPSSL
VFTPFASSSDAFFSGTQQPPQGSVEDEGPKVRVLPERISLQSQEEAEKKKGKIPPPVPKKPSVL
YLPLTSPTAQMEAYVAEPRLPLSPIITLEEDTKCPATGDDLQSLGSSVEPGTEEKSLISDKTAE
WIAEDDDDVFVASRTTEDLFTVIHRSKRKLLGWKEPGSGSSANLDAGRNDDFKALLQKKGSKAT
PRSRPSAAELLKTTNPLARRIIAQFSKDYETTDNPST

NHSL3

MVVFVGRRLPALLGLFKKKGSAKAENDKHLSVGPGQGPGSAVDEHQDNVFFPSGRPPHLEELHT
QAQEGLRSLQHQEKQKLNKGGWDHGDTQSIQSSRTGPDEDNISFCSQTTSYVAESSTAEDALSI
RSEMIQRKGSTFRPHDSFPKSGKSGRRRRERRSTVLGLPQHVQKELGLRNEREAPGTPRAPGAR
DAVRIPTVDGRPRGTSGMGARVSLQALEAEAEAGAETEAMLQRHIDRVYRDDTFVGRSTGTRAP
PLTRPMSLAVPGLTGGAGPAEPLSPAMSISPQATYLSKLIPHAVLPPTVDVVALGRCSLRTLSR
CSLHSASPASVRSLGRFSSVSSPQPRSRHPSSSSDTWSHSQSSDTIVSDGSTLSSKGGSEGQPE
SSTASNSVVPPPQGGSGRGSPSGGSTAEASDTLSIRSSGQLSGRSVSLRKLKRPPPPPRRTHSL
HQRGLAVPDGPLGLPPKPERKQQPQLPRPPTTGGSEGAGAAPCPPNPANSWVPGLSPGGSRRPP
RSPERTLSPSSGYSSQSGTPTLPPKGLAGPPASPGKAQPPKPERVTSLRSPGASVSSSLTSLCS
SSSDPAPSDRSGPQILTPLGDRFVIPPHPKVPAPFSPPPSKPRSPNPAAPALAAPAVVPGPVST
TDASPQSPPTPQTTLTPLQESPVISKDQSPPPSPPPSYHPPPPPTKKPEVVVEAPSASETAEEP
LQDPNWPPPPPPAPEEQDLSMADFPPPEEAFFSVASPEPAGPSGSPELVSSPAASSSSATALQI
QPPGSPDPPPAPPAPAPASSAPGHVAKLPQKEPVGCSKGGGPPREDVGAPLVTPSLLQMVRLRS
VGAPGGAPTPALGPSAPQKPLRRALSGRASPVPAPSSGLHAAVRLKACSLAASEGLSSAQPNGP
PEAEPRPPQSPASTASFIFSKGSRKLQLERPVSPETQADLQRNLVAELRSISEQRPPQAPKKSP
KAPPPVARKPSVGVPPPASPSYPRAEPLTAPPTNGLPHTQDRTKRELAENGGVLQLVGPEEKMG
LPGSDSQKELA

- Appendix B -

Construct name	Amino acids	Domains	Size (kDa) (excluding. N-terminal tag)	N-terminally tagged with....	Size (kDa) (including. N-terminal tag)	Functional assay
NHSC1	1430-1630	Half of putative WH2 domain and the C-terminal epitope for the anti-NHS antibody	24.12 (201 aa)	GST	50.12	Characterisation of a c-terminal antibody Subcellular Localisation
NHSC2	471-1630	CA domains 1 and 2, WH2 domain	139.20 (1160 aa)	Myc	140.52	GST pull-down/actin assay Subcellular Localisation
NHS-CA1	616-704	Putative CA domain 1 + additional a.a. sequence	10.68 (89 aa)	GST	36.68	Actin assay
NHS-CA2	500-615	Putative CA domain 2 + additional a.a. sequence	13.92 (116 aa)	GST	39.92	Actin assay

Table 1 Cloned domains of the NHS protein in prokaryotic and/or mammalian expression vectors.
Refer to figures 6.7 and 6.19 for a schematic representation of NHS construct

- Appendix C -

Publications arising from this study

Identification of the gene for Nance-Horan syndrome (NHS).
J Med Genet. 2004 Oct;41(10):768-71.

Refinement of the X-linked cataract locus (CXN) and gene analysis for CXN and Nance-Horan syndrome (NHS).
Ophthalmic Genet. 2004 Jun;25(2):121-31.

SHORT REPORT

Identification of the gene for Nance-Horan syndrome (NHS)

S P Brooks, N D Ebenezer, S Poopalasundaram, O J Lehmann, A T Moore, A J Hardcastle

J Med Genet 2004;**41**:768–771. doi: 10.1136/jmg.2004.022517

Research report

Taylor&Francis
healthsciences
Ophthalmic Genetics 1381-6810/04/
US$ 22.00

Ophthalmic Genetics – 2004, Vol. 25
No. 2, pp. 121–131
DOI: 10.1080/13816810490514360

Accepted 23 November 2003

Refinement of the X-linked cataract locus (CXN) and gene analysis for CXN and Nance-Horan syndrome (NHS)

Simon P. Brooks[1]
Neil D. Ebenezer[1]
Subathra Poopalasundaram[1]
Eamonn R. Maher[2]
Peter J. Francis[1,3]
Anthony T. Moore[1,3]
Alison J. Hardcastle[1]